Received: 2 June 2020 | Revised: 16 April 2021 | Accepted: 19 April 2021

DOI: 10.1002/esp.5140

RESEARCH ARTICLE

ESPL WILEY

Assessing the hydrological and geomorphic behaviour of a landscape evolution model within a limits-of-acceptability uncertainty analysis framework

Jefferson S. Wong[1] | Jim E. Freer[1,2,3] | Paul D. Bates[1,2] | Jeff Warburton[4] | Tom J. Coulthard[5]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK

[2]The Cabot Institute, University of Bristol, Bristol, UK

[3]University of Saskatchewan Coldwater Laboratory, Canmore, AB, Canada

[4]Department of Geography, Durham University, Durham, UK

[5]Energy and Environment Institute, University of Hull, Hull, UK

Correspondence
Jefferson S. Wong, School of Geographical Sciences, University of Bristol, Bristol, UK.
Email: jefferson.wong@usask.ca

Present address
Jefferson S. Wong, Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK, Canada

Funding information
University of Saskatchewan; Global Water Futures; Royal Society; Environment Agency; University of Bristol

Abstract

Landscape evolution models (LEMs) have the capability to characterize key aspects of geomorphological and hydrological processes. However, their usefulness is hindered by model equifinality and paucity of available calibration data. Estimating uncertainty in the parameter space and resultant model predictions is rarely achieved as this is computationally intensive and the uncertainties inherent in the observed data are large. Therefore, a limits-of-acceptability (LoA) uncertainty analysis approach was adopted in this study to assess the value of uncertain hydrological and geomorphic data. These were used to constrain simulations of catchment responses and to explore the parameter uncertainty in model predictions. We applied this approach to the River Derwent and Cocker catchments in the UK using a LEM CAESAR-Lisflood. Results show that the model was generally able to produce behavioural simulations within the uncertainty limits of the streamflow. Reliability metrics ranged from 24.4% to 41.2% and captured the high-magnitude low-frequency sediment events. Since different sets of behavioural simulations were found across different parts of the catchment, evaluating LEM performance, in quantifying and assessing both at-a-point behaviour and spatial catchment response, remains a challenge. Our results show that evaluating LEMs within uncertainty analyses framework while taking into account the varying quality of different observations constrains behavioural simulations and parameter distributions and is a step towards a full-ensemble uncertainty evaluation of such models. We believe that this approach will have benefits for reflecting uncertainties in flooding events where channel morphological changes are occurring and various diverse (and yet often sparse) data have been collected over such events.

KEYWORDS
CAESAR-Lisflood, GLUE, landscape evolution models, limits of acceptability, observational uncertainty, parameter uncertainty, uncertainty analysis

1 | INTRODUCTION

In most landscapes, processes of weathering, erosion and deposition are highly integrated with hydrological processes and river flow. For several decades, research has tried to unravel the complexity of this behaviour using various methods, increasingly depending on the descriptive and predictive capabilities of numerical models to do this (Owens & Collins, 2006). A part of this

research effort has been focused on developing LEMs to simulate and visualize various geomorphological and hydrological process dynamics simultaneously and examine potential short-term and long-term process linkages. As a result, LEMs are increasingly used in a wide range of applications characterizing various aspects of geomorphological change, such as: (1) testing hypotheses about landform process dynamics (e.g. Densmore et al., 1998; Gioia & Lazzari, 2019; Tucker & Slingerland, 1994); (2) evaluating the effect of changing climate on river morphology (e.g. Coulthard et al., 2000; Coulthard & Macklin, 2001; Hancock, 2009; Hancock et al., 2017; Temme et al., 2009); (3) as a management tool for environmental problems (e.g. Coulthard & Macklin, 2003; Hancock et al., 2000, 2016); and (4) assessing the effects of digital elevation model (DEM) resolution on model simulated outputs (e.g. Claessens et al., 2005; Schoorl et al., 2000) and vegetation effects (e.g. Bastola et al., 2018; Collins et al., 2004; Hooke et al., 2005).

The landscape evolution modelling community has made significant advances in understanding model complexity and component interactions (Van de Wiel et al., 2011), but validation and uncertainty investigations have been limited (Skinner et al., 2018; Tucker & Hancock, 2010). However, to ensure effective use of LEMs, quantifying the magnitude and sources of uncertainty associated with observed constraining data and model simulations is essential as this can increase the reliability of the model predictions and effectively define realistic values that should be used in subsequent assessments. This paper addresses this important issue by exploring the parameter and predictive uncertainty of a LEM by assessing the ability of hydrological and geomorphic uncertain observations to constrain model simulations.

Uncertainty in LEMs is currently acknowledged as an issue, as it is for all environmental modelling, but has rarely been quantified (Skinner et al., 2018; Van de Wiel et al., 2011). Whilst the initial form of the landscape and external driving conditions are often interpolated or extrapolated due to the paucity of available data, the model structure, choice of geomorphic processes, processes formulations and parameterization are strongly affected by a lack of prior knowledge and the difficulty in deciding whether to include or neglect certain processes. These epistemic (knowledge) uncertainties are inherent in the model calculations in terms of cell grid structure and time step (Temme et al., 2011). Whatever the uncertainties, most LEMs have a large number of possible parameters, and each can combine many different ranges of acceptable values. Past studies have mainly focused on LEM sensitivity to changes in climate variability and precipitation characteristic (Armitage et al., 2018; Coulthard & Skinner, 2016; Skinner et al., 2020), and variations in initial conditions (Hancock et al., 2016; Ijjaszvasquez et al., 1992; Kwang & Parker, 2019). Hancock (2009) provided limited evaluation of the sensitivity of model outputs to different perturbations and changes in grain size distributions, whereas Ziliani et al. (2013) examined model sensitivity to 12 input factors as a pre-screening before model calibration was applied. Similarly, Temme et al. (2009) explored different levels of parameter uncertainty and how this affected the ability of the LEM to differentiate between future landscape change under a stable climate and under human-induced climate change. Research conducted by Skinner et al. (2018) explored model sensitivity to parameters and parameter changes of a LEM by using the Morris method for sensitivity analysis. However, a number of questions regarding the effects of LEM uncertainty remain unanswered, for instance, how to identify realistic changes simulated by the LEMs given the parameter and data uncertainties, and how the uncertainty in LEMs propagates when the results of LEMs are used for subsequent analysis. For example, LEM results could be used as inputs in flood inundation models to account for how morphological change impacts on flood risk. By identifying realistic simulations by the LEMs and uncertainty propagation, this is, therefore, a crucial first step in quantifying the parameter uncertainty in LEMs so as to improve their reliability as physically based numerical models.

In common with other types of environmental model applications, LEMs suffer from equifinality (Beven, 1996; Hancock et al., 2016) such that there can be several or many combinations of parameter sets which result in equally acceptable simulations when the model is evaluated against observed data. Numerous approaches (e.g. standard Bayesian approaches (Krzysztofowicz, 2002; Kuczera & Parent, 1998), Bayesian model averaging (BMA) (Ajami et al., 2007; Duan et al., 2007; Vrugt et al., 2006), generalized likelihood uncertainty estimation (GLUE) (Beven & Binley, 1992) and others) have been proposed on the basis of accepting multiple acceptable parameter sets and treating each as a scenario of uncertain conditions that describes a simulated system. In most of the real applications, since a residual time series is neither available nor independent of uncertain observed data (Winsemius et al., 2009), the justification for making strong statistical assumptions about the nature of likelihood functions is rather weak. It will be difficult to apply formal statistical parameter inference which involves updating a prior distribution of model parameters based on statistical likelihood measures and requires an explicit account of errors in the model structure, parameters and the input data (Mantovan & Todini, 2006). Given the fact that the residual error characteristics are not fully known and the issues with a lack of commensurability between the limited observed data and the model, GLUE is a more suitable approach to reveal uncertainties in LEMs from which a probabilistic prediction can be made at the expense of relaxing certain statistical assumptions of the formal Bayesian approach (Beven, 2009).

Since there is no objective method to choose the threshold for some informal likelihood measures (Mantovan & Todini, 2006; Montanari, 2005), GLUE has been criticized for its subjective distinction between behavioural and non-behavioural model. In response, an approach to model evaluation on the basis of limits of acceptability (LoA) for use within the GLUE methodology was outlined by Beven (2006) and first demonstrated by Liu et al. (2009). This approach suggested that, before simulating any model runs, the LoA should first be defined based on a range of 'effective observational error' that incorporates observation error in the measurements given the availability of information and allows for the effects of commensurability error and input error. The Monte Carlo model runs that provide predictions that all fall within the LoA are then classified as behavioural, and each model is associated with a performance score that summarizes how well the model simulates the observed. Since different error uncertainties are often inherent within different types of data, interpretation of the model results could be biased because

the errors from various sources (i.e. input data, model structure, model parameters and observations for calibration and validation) could compensate each other (McMillan et al., 2012). The LoA approach thus pays attention to different sources of uncertainty and allows different limits to be set for individual observations in the calibration process such that the modelling exercise is fit for purpose. Liu et al. (2009) demonstrated the use of the LoA approach in GLUE for identifying behavioural models whilst allowing for uncertainties in observational data. Other studies which focused on the calibration of hydrological models have adopted this approach in building their LoA for: (i) uncertainties in the stage–discharge relationship and evaluation points in a flow–duration curve (Van Hoey et al., 2015; Westerberg et al., 2011b); (ii) flood frequency estimation (Blazkova & Beven, 2009); (iii) evaluation of model structure, parameter and data uncertainties (Krueger et al., 2010; Mackay et al., 2018; Teweldebrhan et al., 2018); (iv) calibration of models incorporating both hard and soft information (Winsemius et al., 2009); and (v) uncertainties in hydrological and water quality data (Coxon et al., 2015; Hollaway et al., 2018; McMillan et al., 2012). It is therefore expected that this approach will be equally applicable in other environmental modelling frameworks such as LEMs.

The aim of this paper is to apply an LoA uncertainty analysis approach to evaluate LEM simulations: (1) to quantify and assess the value of uncertain hydrological and geomorphic data in constraining the catchment response, and (2) to identify the uncertainty of parameters for predictions of observed and uncertain hydrological and geomorphic behaviours and dynamics. The River Derwent and Cocker catchment in North West England, UK provides a good test of the methodology because of the availability of stage–discharge and suspended sediment load data, which are both highly uncertain and so require the quantification of the expected observational errors (see Table 1 for data summary). Also, the locations of the gauges and the monitoring sites provide an opportunity to assess not only the model performance at the catchment outlet but also the model ability in capturing the internal catchment response. Such an evaluation is uncommon in most LEM applications.

2 | STUDY AREA AND DATA AVAILABILITY

2.1 | Catchment characteristics

Located in North West England, the River Derwent catchment covers an area of 663 km^2 (Environment Agency, 2009) (Figure 1). The River Derwent rises in the high peaks of the Lake District, flows through Derwent Water and Keswick and continues into Bassenthwaite Lake and down to Cockermouth before draining into the Irish Sea at Workington. Key tributaries include the River Cocker (catchment area 117 km^2), which rises at Gatesgarthdale Beck, flows through the lakes of Buttermere and Crummock Water and joins the River Derwent at Cockermouth. The River Greta, formed by St Johns Beck and the River Glenderamackin, has its confluence with the River Derwent just downstream of Keswick. The combined catchment area for the River Greta and St Johns Beck is 143 km^2.

The catchment is very steep in its upstream sections, which contain some of the highest peaks in England (over 900 m). The catchment geology is dominated by the Skiddaw Slate Group, which the River Cocker and most of the River Derwent upstream of Cockermouth (including Bassenthwaite Lake and Derwent Water) flow over, while the upper reaches of the River Derwent, Naddle Beck and St Johns Beck lie on the Borrowdale Volcanic Group (Hatfield & Maher, 2008, 2009). The remainder of the Derwent lies on Carboniferous limestone, milestone grit and coal measures (Moseley, 1979; Wilson, 2010). Holocene (post-glacial) alluvium (river sediments) occur along many of the main valleys in the catchment, which are dominated by loamy soils. The watercourses within the catchment comprise steep bedrock channels with step-pool sequences in the headwaters and boulder/gravel-bed channels in valley reaches. Sediment sizes range from sand to boulders but are dominated by gravel and cobbles. Vegetation within the catchment is dominated by grassland (Environment Agency, 2009). The average annual rainfall is 2,408 mm but can be as high as 4,175 mm on the mountaintops (Environment Agency, 2009). In the upper headwaters, the rivers have a very flashy flow regime due to topography, geology and soils. However, this flashy response of the upstream reaches is attenuated by the lakes in the downstream sections of the River Derwent (Hatfield & Maher, 2008). The River Cocker has a more flashy response than the River Derwent. Combined with the impermeable underlying geology and waterlogged upland soils, large amounts of runoff are produced, and this can cause significant downstream flooding (Chiverrell et al., 2019).

2.2 | Data

To improve the understanding of the catchment dynamics, a NEXTMap DEM at 50 m resolution of the entire catchment was used. This DEM was provided by Intermap Ltd. based on an airborne interferometric synthetic-aperture radar (IFSAR) survey with a vertical accuracy of 1 m. Rainfall data were provided by the Environmental Agency of England and Wales (EA) during the period from December 1990 to September 2012. To account for the spatial and altitudinal effect, rainfall data from 21 tipping-bucket rain gauges located within and 6 km around the catchment (see Figure 1 for spatial distribution) were obtained to produce an areal average rainfall (AAR) for both the River Derwent and Cocker catchments. Stage–discharge measurements at five flow gauging stations (Figure 1 and Table 1) were obtained from the EA, with two main gauges (Derwent catchment at Ouse Bridge and Cocker catchment at Southwaite Bridge) and three sub-catchment gauges (Derwent at Portinscale, Glenderamackin at Threlkeld and St Johns Beck at Thirlmere Reservoir) in the River Derwent catchment.

For suspended sediment, the data reported in the study of Warburton (2010) in which fluvial sediment flux monitoring was undertaken in the Derwent catchment from 1 April to 30 November 2006 were used (Table 1). River stage (m) and turbidity (measured in nephelometric turbidity units using an Analite 395 nephelometric turbidity probe) were recorded continuously and logged at 15-min intervals on a Campbell CR200 data logger at two monitoring sites (Derwent at Portinscale and Newland Beck at Braithwaite). The 15-min interval data were re-grouped to hourly time intervals and processed by Warburton (2010) to provide the suspended sediment data. While no primary field data were collected during this study, the grain size distribution (GSD) characteristics of the Eden catchment, a

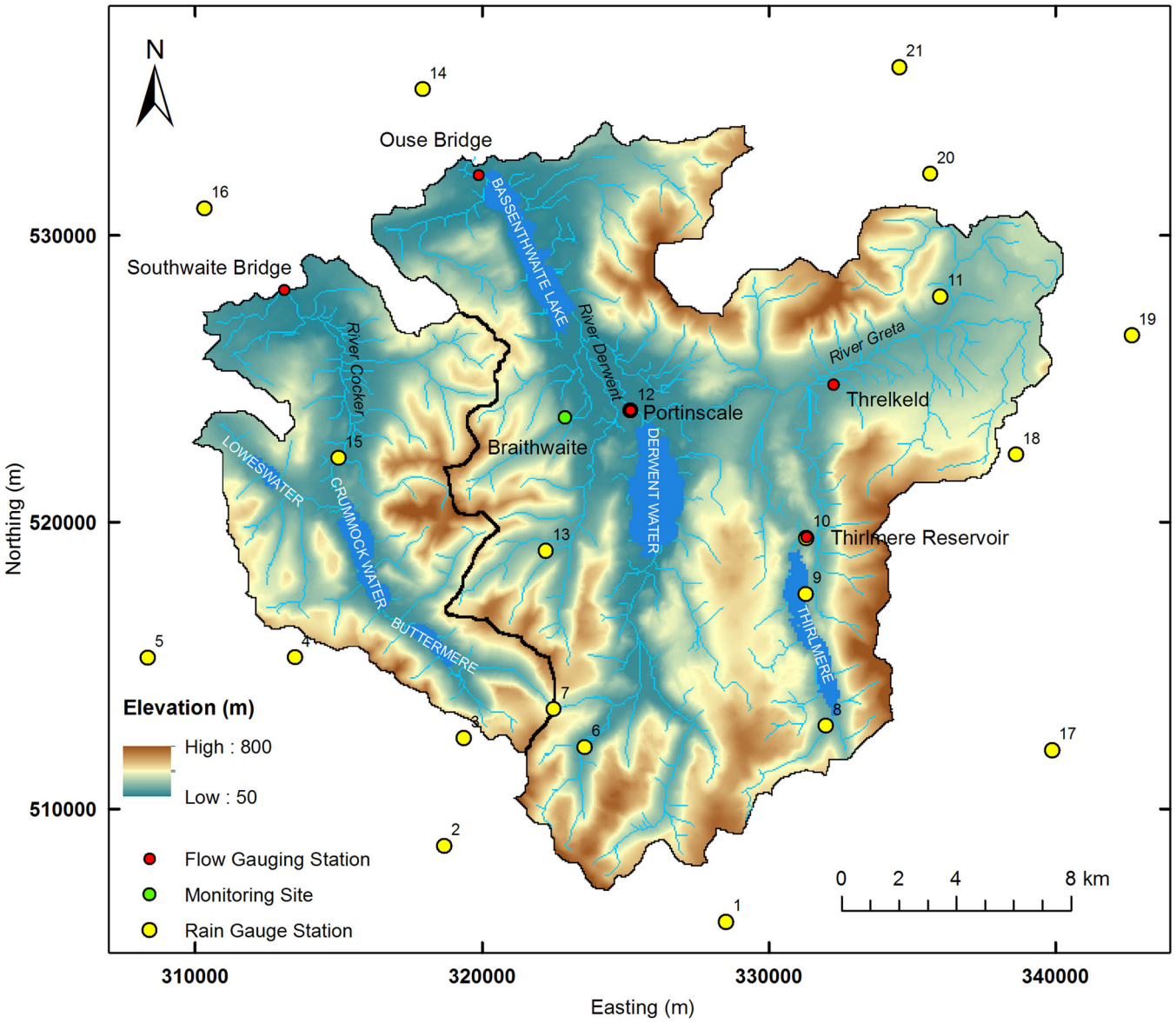

FIGURE 1 DEM of River Derwent and Cocker catchment extracted from LiDAR data at 50 m resolution. Red dots are gauging station points: River Derwent at Ouse Bridge for River Derwent catchment and River Cocker at Southwaite Bridge for Cocker catchment. Yellow points indicate the rain gauges within and 6 km around the catchment. Number of each point is numerical code of the gauge (refer to Table 1 for details)

TABLE 1 Summary of the available hydrological and geomorphic information. The rainfall and flow data are provided by EA

Variable	Description	Data source	Number of stations	Observation interval	Start	End
Rainfall	Tipping-bucket	EA	21	15 min	December 1990[a]	September 2012[a]
Flow	Stage and discharge	EA	5	15 min	June 1966[a]	May 2012[a]
Suspended sediment	Stage and turbidity	Warburton (2010)	2	15 min	April 2006	November 2006

[a]The start and end times represent the earliest and latest dates, respectively, for the rain gauge and the flow gauging stations.

neighbouring catchment north east to the River Derwent and Cocker, were used as the input of the model in this study. The Eden catchment was a suitable substitute because: (1) it shared similar upland setting; (2) Holocene alluvium is present along much of the main valleys in the catchment; (3) catchment soils, particularly those in river valleys, are composed principally of loams with different percentage of clay content; and (4) rivers are dominated by gravels and pebbles with occasional boulders and bedrock outcrops. The grain size distributions of the six sub-catchments of the Eden were estimated using a photo analysis technique in which 173 photographs were taken on channel edge at 40 sites along the rivers (personal communication with Jorge Ramirez).

2.3 | Pre-processing

The DEM was first rescaled from 50 m to 200 m and 100 m spatial resolutions using bi-linear interpolation; there are 8,916 and 2,945 grid cells for the River Derwent and Cocker catchments, respectively. The purposes of rescaling are to increase the computational efficiency and enable multiple simulations to be run. Summary statistics of the elevation and slope of the DEMs at 50 m, 100 m and 200 m resolutions were computed and their comparison showed that the elevation and slope at coarser resolutions provided similar values at different percentiles to those at 50 m resolution (see Table A1 and A2). Therefore, the 200 m and 100 m DEMs were able to capture the dominant topographic features and provide enough detail to reflect the dominant catchment-scale topographic spatial heterogeneities with a feasible simulation time. Given the presence of six lakes in the catchment and the fact that the DEM could only capture their water surface elevation, the DEM was modified to provide a better representation of the lake topography. The bottom topography of four of the lakes (Bassenthwaite Lake, Derwent Water, Buttermere and Crummock Water) was based on the bathymetrical surveys of Mill (1895). Mill's survey maps were geo-referenced (with an average root-mean-square error (RMSE) of 5.30 m), and the contour lines were digitized manually, and converted into 100 m and 200 m raster DEMs using the ArcMap 10.1 raster interpolation and the kriging algorithm. The resulting DEMs were integrated into the NEXTMap 100 m and 200 m DEMs by subtracting the lake depths from the water surface elevation. Since no survey data were available in literature for the remaining two lakes (Thirlmere and Loweswater), their bathymetry was configured using their mean depths such that the model lake storage matched the actual capacity measured by the EA (Environment Agency, 2006).

The 15-min interval rainfall data were first re-grouped to hourly time intervals before screening. The following criteria were applied to the screening procedure: (1) compute the percentage of missing values in each of the rainfall series on a yearly basis and eliminate any years for which the percentage of missing values exceeds 10%; (2) maximize the numbers of years so that the rainfall data series are of reasonable length, ca. 10 years; (3) retain as many rain gauges within the catchment as possible even though their percentage of missing values in some years exceeds 10%; and (4) evaluate the rain gauge consistency against the three nearest neighbours by correlation analysis. Accordingly, the hourly rainfall series of the rain gauges (except rain gauges 1, 18, 19 and 20, Figure 1) from 1999 to 2011 were used and regarded as the observed rainfall period. The hourly rainfall data of the 17 rain gauges were interpolated using the ordinary kriging method, which preserves the pattern of spatial dependency (see e.g. Goovaerts, 2000; Mair & Fares, 2011) to estimate the spatial precipitation field of both catchments at an hourly time step. The total gridded spatial rainfall within the catchment was divided by the catchment area to produce the hourly homogeneous AAR. This was done separately for the River Derwent and River Cocker catchments. To assess the robustness of the time series, the AARs of both catchments were aggregated into daily values, and its spatial field and temporal variability were compared with daily 1 km-gridded rainfall data, which is a composite of radar data and rain gauge data provided by the EA. In general, the AARs showed similar rainfall characteristics (with coefficient of determination, R^2, of 0.61 and 0.57 for the River Derwent and River Cocker catchments, respectively) and were compatible with daily 1 km-gridded rainfall data.

3 | THE CAESAR-LISFLOOD MODEL

CAESAR-Lisflood is an LEM that simulates the evolution of landforms by directing water over a regular grid of cells and modifying elevations based on erosion and deposition caused by fluvial and slope processes (Coulthard et al., 2013). CAESAR-Lisflood integrates the LISFLOOD-FP 2D hydrodynamic flow model (Bates et al., 2010) with the CAESAR geomorphic model (Coulthard et al., 2000, 2002, 2005; Coulthard & Van De Wiel, 2007; Van De Wiel et al., 2007) to dynamically simulate both erosion and deposition and flood inundation extent and depth simultaneously in river catchments and reaches over a range of temporal scales. There are four main components in CAESAR-Lisflood, featuring hydrological processes, multidirectional routing of river flow, fluvial erosion and deposition over a range of different grain sizes, and slope processes (soil creep, mass movement). These components are described briefly below, but for a full description see Coulthard et al. (2002), Van De Wiel et al. (2007) and Coulthard et al. (2013).

CAESAR-Lisflood can be run in two modes: a catchment mode, with no external influxes other than rainfall; and a reach mode, with one or more points where discharge and sediment enter the system. When running in catchment mode, hourly rainfall data are used to drive an adapted version of TOPMODEL (Beven & Kirkby, 1979) to calculate runoff, which is then routed using the flow model. In reach mode, sources of discharge (both water and sediment) can be added at user-defined points. Surface water routing is then carried out using the 2D hydrodynamic LISFLOOD-FP model (Bates et al., 2010). The hydraulic model time step is controlled by the shallow-water Courant–Friedrichs–Lewy condition to maintain numerical stability. The flow depth and flow velocity are used to calculate shear stress, which in turn determines the erosion, transport and deposition of sediment. CAESAR-Lisflood can simulate erosion and deposition over nine sediment fractions, with one fraction treated as suspended sediment. As the study channels are largely gravel and sand, sediment transport is calculated using the Wilcock and Crowe (2003) equations, which are based on field and laboratory data from a coarser bed gravel–sand mix. Deposition of sediments differs between bed load and suspended load, with bed load being moved directly from cell to cell, whereas suspended load is deposited according to fall velocities and concentrations for the suspended sediment fraction. Erosion within the channel is controlled by the in-channel lateral erosion rate, which represents how cohesive or not the sediment is.

Slope processes are also included, with mass movement and soil creep. Mass movement (landslides) is represented as an instantaneous movement process. When a critical slope threshold is exceeded by the slope between adjacent cells, material is moved from the uphill cell to the one below until the angle is lower than the threshold. Movement upslope may be triggered by a small slide in a cell at the base of the slope, and the adjacent cells are checked iteratively until there is no more movement (Coulthard et al., 2002). Soil creep (per year) is calculated between each cell using a simple diffusion equation which is linearly proportional to slope angle. After the fluvial erosion/

deposition and slope process amounts are calculated, the elevations and grain-size properties of the cells are updated simultaneously. In this study, CAESAR-Lisflood version 1.2x was used.

4 | METHODOLOGY

4.1 | Model setup

CAESAR-Lisflood was set up in catchment mode, and its main data sources were a DEM as the landscape, hourly rainfall inputs and grain size distribution (GSD). The river channel and floodplain and lake topography in the model were described using the modified DEM as above (i.e. 200 m and 100 m DEM for the River Derwent and Cocker catchments, respectively). Bedrock can afford an important control on channel incision – especially in upland areas and over long (e.g. millennial) timescales. However, mapping the bedrock depth and any locations where it is at the surface is a considerable task for such a large catchment. Furthermore, for this study, many parts of the simulated basin were low gradient and lacustrine, where the impact of bedrock will be less than in an entirely upland basin. Therefore, the simulations were run without a bedrock layer present.

Many numerical models require a spin-up period for the results to become stable. In the case of CAESAR-Lisflood, the model produces extremely high sediment transport rates in early simulations as surface roughness in the digital elevation is removed and smoothed and the particle size distribution is sorted across the catchment according to the topography and hydrology. In this case, the model was usually run using repeated rainfall data for years of simulation (e.g. Hancock (2009) allowed two cycles of the rainfall data) and the resultant DEM and GSD data were then used for analysis. To avoid the repeated stochastic nature of the rainfall series, the 13-year hourly AAR series was extended to reproduce a new series by using the cumulative density function and generalized Pareto distribution model, a stochastic rainfall generator as described in Cameron et al. (1999). The generated rainfall series was added in front of the observed AAR to form a 26-year-long time series extending from January 1986 to December 2011. The years 1986–1998 were used as a spin-up period for the model, thus no model evaluations were made in this period. The 26-year hourly AAR time series was then applied uniformly across each catchment.

Since the river channels are dominated by gravel and cobbles with sediment sizes ranging from sand to boulders, it is important to reflect the sediment variability of the catchment in order to examine the effects of GSD on model performance and the hydrological and geomorphic behaviour of the model during floods. In this regard, the grain size data were classified into nine size ranges to suit CAESAR-Lisflood (Table 2). Since grain sizes smaller than 0.3 mm were not observable using the photo analysis technique, the GSDs were adjusted by assuming a 20% proportion of fine sediments for each sub-catchment (personal communication with Jorge Ramirez). Figure 2 shows the uncertainty bound (grey) from 0.5 mm ($+1\phi$) to 128 mm (-7ϕ) based on the variability (5th and 95th percentile) of the field measurements of all sub-catchments, while the envelope less than 0.5 mm is typical of upland soils described in Wilson (1993). The black line (case 1 in Table 2) represents the mean GSD characteristic of all sub-catchments. To provide a basis for varying the GSD, the mean GSD was first described by four grain-size parameters which are based on Folk and Ward (1957) method: (a) the average size (mean), (b) the spread (sorting) of the sizes around the mean, (c) the symmetry or preferential spread (skewness) to one side of the mean and (d) the degree of concentration of the grains relative to the mean (kurtosis). Only variations in mean and kurtosis were of interest in this study, thus cases 2 to 5 (Figure 2 and Table 2) were selected by changing the mean and kurtosis of the distribution while maintaining the skewness and sorting of the distribution to fall into the same description. One grain size parameter is changed in each case to further reduce the numbers of cases that need to be simulated. In general, case 2 can be described as 'peaky', case 3 as 'less peaky', case 4 as 'coarser' and case 5 as 'finer', relative to case 1, to reflect different distribution characteristics.

4.2 | Model parameters and sampling range

There are a total of 24 parameters in the CAESAR-Lisflood model that must be specified and are normally treated as being homogeneous across the model domain. To reduce the dimensionality of the

TABLE 2 Grain size proportions in different cases

	Case 1	Case 2	Case 3	Case 4	Case 5
Adjusted size (mm)	Proportion (%)				
0.063	10.0	7.3	12.0	6.2	12.3
0.25	10.0	8.7	12.1	8.8	12.5
1	12.0	9.7	12.5	6.4	17.9
2	24.0	28.2	19.4	19.4	26.1
4	21.0	27.0	18.2	25.1	18.6
8	13.0	11.6	17.5	18.2	7.5
16	6.0	5.0	7.4	10.0	2.6
32	2.0	1.5	0.4	3.5	1.9
128	2.0	1.0	0.5	2.4	0.6

Case 1 represents the mean grain size distribution characteristic of all sub-catchments of the Eden. Cases 2 to 5 are selected by changing the skewness and kurtosis of the distribution while maintaining the mean and sorting of the distribution to fall into the same description. One grain size parameter is changed in each case. In general, case 2 can be described as 'peaky', case 3 as 'less peaky', case 4 as 'coarser' and case 5 as 'finer', relative to case 1

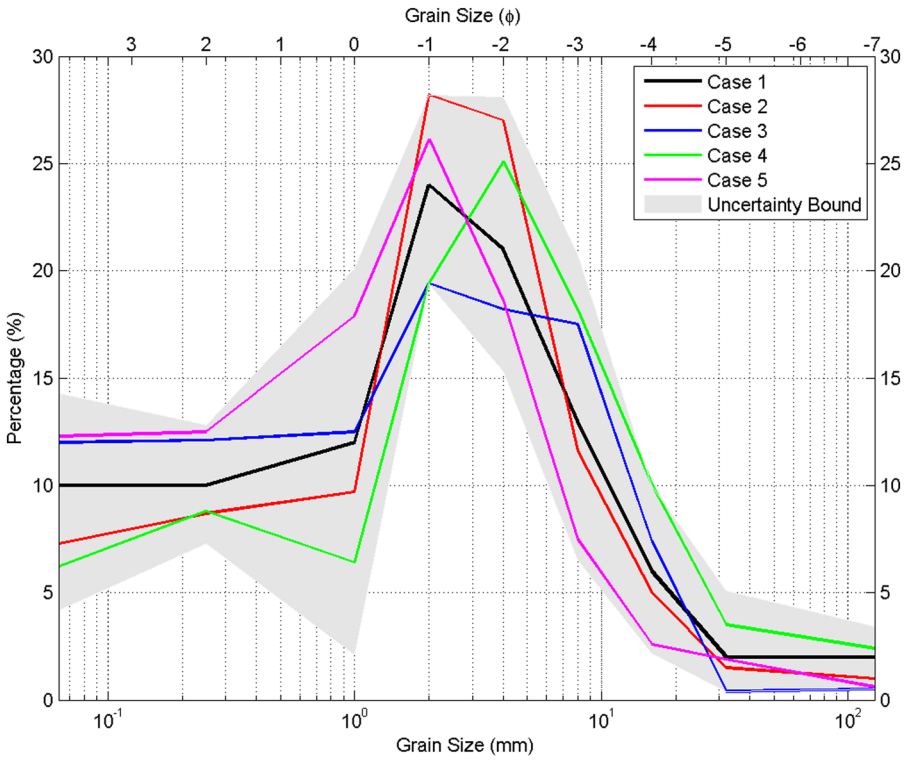

FIGURE 2 Grain size distribution used for input into CAESAR-Lisflood. The uncertainty bound (grey) from 0.5 mm (+1ϕ) to 128 mm (−7ϕ) is based on the variability of the field measurements of all sub-catchments, while the envelope less than 0.5 mm is indicated by the ranges described in Wilson (1993). Case 1 (black line) represents the mean grain size distribution characteristic of all sub-catchments of the Eden. Cases 2 to 5 are selected by changing the mean and kurtosis of the distribution while maintaining the skewness and sorting of the distribution to fall into the same description. One grain size parameter is changed in each case. In general, case 2 can be described as 'peaky', case 3 as 'less peaky', case 4 as 'coarser' and case 5 as 'finer', relative to case 1

parameter space, five parameters of the model were kept at their 'default' values due to the model's lack of sensitivity to variations in these factors in previous studies (Skinner et al., 2018; Ziliani et al., 2013). Also, five parameters (two for bedrock lowering and three for physical weathering) in the soil development component of the model were not included to further reduce the model complexity. We acknowledge that the bedrock might have an impact on the model simulations, but since the focus of the current modelling is on the hydrological and geomorphic behaviour of the model during floods, we assume that the bedrock lowering and physical weathering play a minor role as compared with other factors. As a result, this study identified 14 parameters, which include one hydrologic parameter, three hydraulic parameters, five sediment parameters, two slope parameters and three vegetation parameters, plus the GSD cases as one of the parameters sampled within the GLUE analysis (Table 3). We assumed a priori that parameter distributions were uniform due to a lack of evidence of what the effective model parameter distributions should be. This means that *posterior* parameter distributions are constrained by quantifying model performance to uncertain observed data using the LoA criteria (see Section 4.3 for details).

To improve the sampling of such a high-dimensional parameter space in computationally expensive models such as LEMs, Latin hypercube sampling was employed. This technique ensures that Monte Carlo samples more efficiently cover the parameter interactions in *n* dimensions for a given sparse sample size and so generate minimally correlated parameter sets (Beven & Freer, 2001). Accordingly, a sparse population of 1,500 parameter sets was generated for each catchment given the computational burdens that each simulation took 6 days to 2 weeks to complete (the variability being due to the model dynamics under different parameterizations taking considerably different simulation run times). The ranges of these parameters are assigned based on their physically feasible ranges and from literature values (Table 3). If no values were reported in literature, a default value was set based on expert knowledge and allowed to vary either by ±50% from the default or by a range suggested by previous modelling experience. Alternatively, parameters were allowed to vary within the 95% confidence intervals obtained from observations if available. The practice of specifying the ranges of parameter values by a certain percentage has been seen in previous studies (e.g. Pappenberger et al., 2007) and therefore has some precedence. A summary of these parameters and their sampling ranges is given in Table 3.

4.3 | Quantification of observational error

The starting point for setting LoA is to assess the uncertainty in the observed data that are being used for model evaluation. Uncertainties in input data (e.g. rainfall) could certainly be included to define a set of ensemble simulations. However, making an assessment of input error is limited by the simple representation of the rainfall in the model which could not fully account for the spatial aspects of the rainfall uncertainty. This study therefore focused on the errors in discharge and suspended sediment data arising from uncertainty in the stage–discharge rating curve and sediment load duration curve respectively, where clear evidence was available to quantify these data.

4.3.1 | Stage–discharge rating curve and observational uncertainty

Numerous methods (e.g. log–log linear regressions (Liu et al., 2009), envelope curves (Krueger et al., 2010; McMillan et al., 2010), Bayesian statistics (Moyeed & Clarke, 2005) and fuzzy rating curves (Pappenberger et al., 2006; Westerberg et al., 2011a)) have previously been used to estimate discharge rating curve errors (Kiang et al., 2018). Applying the most commonly used power function

TABLE 3 Parameters, description and ranges used in the GLUE LoA uncertainty analysis

Component	Parameter (unit)	Description	Sample range	Distribution	Reference/source
Hydrological processes	*m (−-)*	A modification of the vegetation parameter used in TOPMODEL	0.005–0.020	Uniform	Welsh *et al.* (2009)
Hydraulic processes	*WD_T (m)*	Water depth threshold over which erosion will happen	±100% from default (0.01 m)	Uniform	Expert knowledge and modelling experience
	DSS (m)	Slope for edge cells	Derwent: 0.0012 m – 0.0019 m Cocker: 0.00067 m – 0.0016 m	Uniform	95% confidence intervals of the best-fit regression line of the bed elevation against a subsection of river to river outlet of the domain
	n (−)	Manning friction coefficient	0.025–0.04	Uniform	Chow (1959)
Sediment processes	*Max_Ero (−)*	Maximum amount of material that can be eroded or deposited within a cell	0.005–0.02	Uniform	Ziliani *et al.* (2013)
	InCh_LatEro_R (−)	In channel lateral erosion rate that controls how cohesive or not the sediment is	5–20	Uniform	Expert knowledge and modelling experience
	LatEro_R (−)	Lateral erosion rate that controls bank failure	0.000001–0.0001	Uniform	Expert knowledge and modelling experience
	Num_Edge_Sf (−)	Number of passes for edge smoothing filter	Derwent: 20–49 Cocker: 15–62	Uniform	5 and 95 percentiles of the numbers of cells counted along the main river based on 50 m DEM
	Num_Cell_Led (−)	Number of cells to shift lateral erosion downstream	Derwent: 2–5 Cocker: 1–6	Uniform	Expert knowledge: 10% of the values used for *Num_Edge_Sf*
Slope processes	*Creep_R (year)*	Creep rate	±50% from default (0.0025 year)	Uniform	Hancock *et al.* (2000, 2002), Hancock (2009)
	Slopefail_T (°)	Slope failure threshold above which landslides happen	Derwent: 30° – 36° Cocker: 40° – 47°	Uniform	Warburton et al. (2008)
Vegetation	*GrassGrow_R (year)*	The speed in which vegetation reaches full maturity in years	0.5 – 3 year	Uniform	Skinner *et al.* (2018)
	Veg_CriShear (−)	A parameter above which vegetation will be removed by fluvial erosion	1–20	Logarithmic	Collins et al. (2004)
	Prop_Ero (−)	Proportion of erosion that can occur when vegetation grown	0–1	Uniform	Expert knowledge and modelling experience

relationships in the River Derwent and Cocker catchments did not give satisfactory results because the assumptions of normality and homoscedasticity of the residuals are not met, thus a non-parametric framework is adopted in this study, similar to the method of Coxon et al. (2015).

The stage–discharge gauging data and observed uncertainty ranges estimated were fitted with the locally weighted scattering smoothing (LOWESS) method of Cleveland (1979). The LOWESS method provides an objective and empirical approach to curve estimation and associated uncertainties that requires no a priori assumption as to the form of the relationship. It is preferred in cases, such as those in this study, where the log–log relationship is non-linear and exhibits curvature (e.g. Hicks et al., 2000). The stage–discharge gauging data were first transformed to obtain a linear relationship through a log transform of gauge-height data and a Box–Cox transform of discharge data, corresponding to a more empirical form of the usually log-transformed stage–discharge power-law function (Moyeed & Clarke, 2005). The lambda parameter in the Box–Cox transformation was optimized to achieve the highest degree of linearity. The transformed stage–discharge data were then fit with LOWESS, where the LoA uncertainty bounds are computed from how well the estimated curve fits the population of stage–discharge gaugings.

Since the maximum stages in the observation period (1999–2011) are higher than those in the historical records at all gauging stations, extrapolation from the above LOWESS curves and uncertainty estimate fittings were needed. This was quantified by assuming a linear relationship from the upper tail of the LOWESS fittings with the same level of uncertainties in the log–log-transformed space. Whilst we recognize that such uncertainties might become relatively larger as discharge magnitudes are increased, we have no evidence to justify a different approach, so this simple extension approach was used. The extended LOWESS rating curves (Figure 3) were then used for the estimate of discharge in the observation period, and the resulting uncertainty bounds consequently defined the maximum and minimum discharge intervals for given stages. As a result, the percentages of the total time series in the observation period that were extrapolated in this form ranged from 3.5% to 17.5% for different gauging stations (Figure 3).

4.3.2 | Sediment load–duration curve

Similar to the stage–discharge measurements, a log transform of suspended sediment concentration and discharge data (Q–C

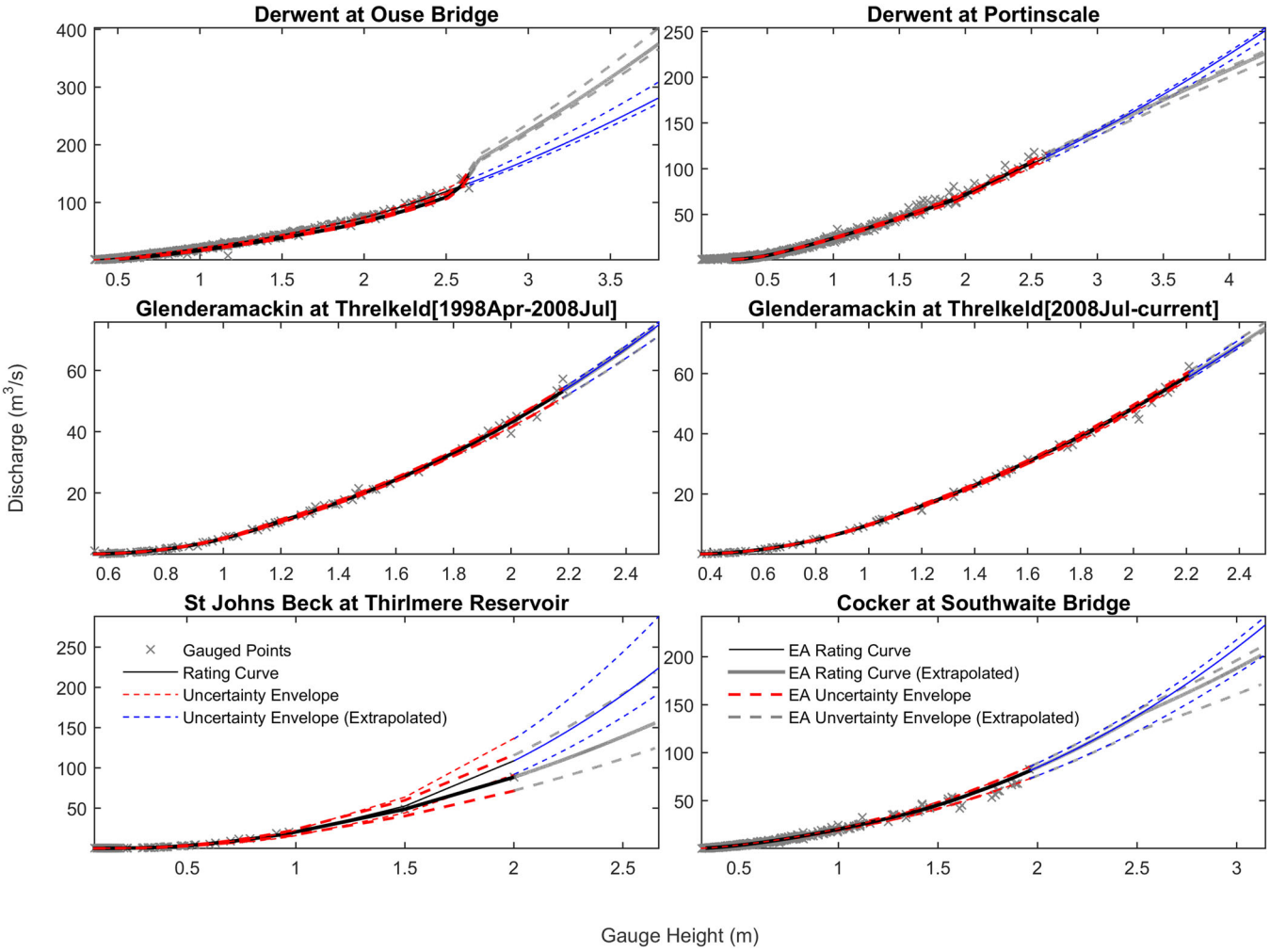

FIGURE 3 Uncertain rating curves for the five flow gauging stations in the River Derwent and Cocker catchment derived from the stage–discharge measurements. The black crosses represent the measured values, and the black solid lines indicate the fitted rating curve, while red and blue dashed lines represent the uncertainty limits for the fitted rating curve and the extrapolated part, respectively. The rating curve from the EA is also plotted in each flow gauging station for reference

relationship) was first applied to find a linearized relationship, but this proved negative. This could be due to a number of well-documented reasons, for instance hysteresis effects, seasonal effects, antecedent conditions in the catchment and temporal change in vegetation cover (Asselman, 2000; Ferguson, 1987; Walling, 1977). In this regard, the sediment load–duration curve was developed to estimate the errors in the suspended sediment data. This adapts the concept of the flow–duration curve that is commonly applied in hydrology and engineering fields, but in this case it indicates the percentage of time that given suspended sediment loads are likely to be exceeded. Suspended sediment loads (tonnes) were calculated by multiplying discharge data by suspended sediment concentrations. Since the sediment outputs in CAESAR-Lisflood were in m^3, a further unit conversion was undertaken from tonnes to cubic metres by assuming the dry bulk density at source to be 1.3 tonnes/m^3 (Verstraeten & Poesen, 2001). Exceedance percentages from the sorted suspended sediment load values were then calculated based on the percentile values 100(0.5/n), 100(1.5/n),..., 100([$n-0.5$]/n), where n is the number of suspended sediment load values.

The accuracy of the sediment load–duration curve was limited by the availability and representativeness of the data, hence a bootstrap technique (Efron, 1979) was employed here to account for the uncertainty of the sediment load–duration curve induced by the concentration data. A total of 10,000 bootstrap samples were taken in this study, and the 95% confidence intervals of the set of samples defined the upper and lower observed LoA for the sediment load–duration curve. In general, the uncertainty bounds were the largest at the high-load part of the sediment load duration curve but were progressively smaller towards low load (Figure 4).

4.4 | Evaluation of model performance

Time-step-based performance measures (Krueger et al., 2010) were used as a means of evaluating the model's hydrological behaviour to reproduce the timing and magnitude of hourly flood events and the model's geomorphic behaviour to replicate the cumulative suspended sediment load. To focus on the catchments' behaviour in producing flood events, the model performance to flood peaks for behavioural simulations was tested against 14 flood events (10 flood events for Glenderamackin at Threlkeld and St Johns Beck at Thirlmere Reservoir sites), with upper and lower limits of acceptability per time step during the flood event periods being given by the observed discharge uncertainty interval outline in section 4.3. The flood events included the six major flood events (December 2003, January 2005, October 2005, December 2006, October 2008 and November 2009) and the annual maximum flood events for the remaining years.

Similarly, to cover the catchments' behaviour in generating suspended sediment loads, the time steps of interest would be the evaluation points from the sediment load–duration curve. Since the high-load part of the sediment load–duration curve contains most of the information about the dynamic response of the catchment to the effective discharge events, points chosen were equally spaced by magnitude rather than equal fractions of total time. This was achieved by using an approach similar to that applied in Westerberg et al. (2011b), where the suspended sediment load values, rather than exceedance percentages, were divided into N equal classes, with $N = 21$ intervals being used for this study (see Figure 4). The maximum and minimum values of the entire sediment load–duration curve were excluded, and the remaining $N-1$ suspended sediment class boundary values were used to identify the corresponding 20 evaluation points. The evaluation points in terms of exceedance percentages for the suspended sediment class boundary values were calculated by linear interpolation between the sorted suspended sediment load values.

A scaled score (S) was calculated to define the deviation of the simulated results from both the observed discharge and sediment load data. The scaled score was calculated relative to the observed uncertainty from the LoA at each evaluated time step, given as

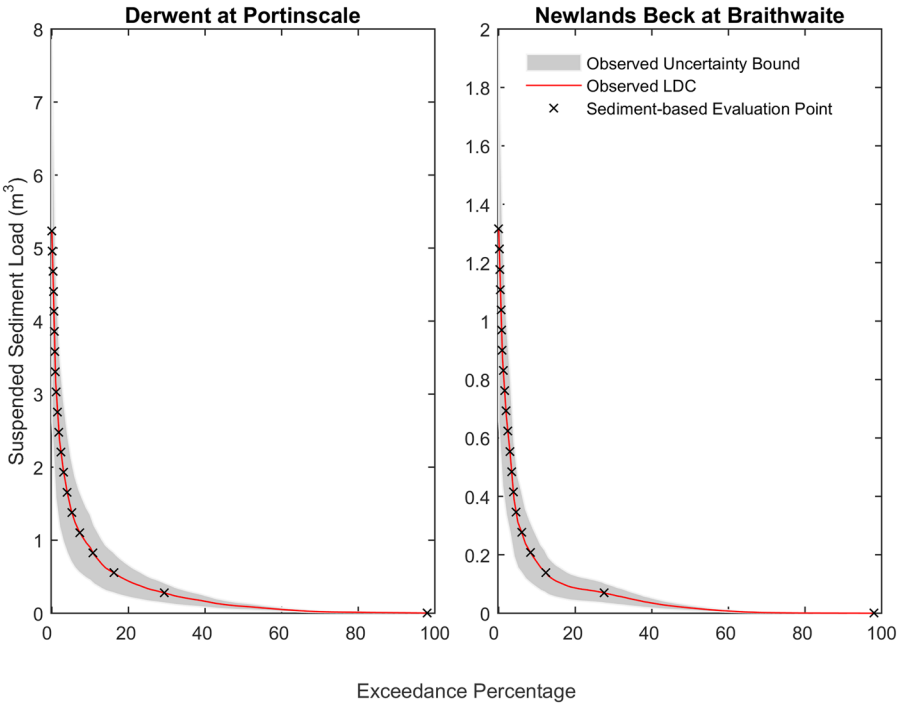

FIGURE 4 Sediment load–duration curve (LDC) for Derwent at Portinscale (left) and Newlands Beck at Braithwaite (right). The black solid crosses indicate the evaluation points using equal intervals of suspended sediment load. The lower and upper acceptability limits at these evaluation points were determined by using a bootstrap resampling of the observed suspended sediment load data and the exceedance percentages

$$S_Q(t) = \begin{cases} (Q_{sim,t} - Q_{obs,t})/(Q_{obs,t} - Q_{min,t}) \text{ if } Q_{sim,t} < Q_{obs,t} \\ (Q_{sim,t} - Q_{obs,t})/(Q_{max,t} - Q_{obs,t}) \text{ if } Q_{sim,t} \geq Q_{obs,t} \end{cases} \quad (7)$$

$$S_{SS}(t) = \begin{cases} (SS_{sim,t} - SS_{obs,t})/(SS_{obs,t} - SS_{min,t}) & \text{if } SS_{sim,t} < SS_{obs,t} \\ (SS_{sim,t} - SS_{obs,t})/(SS_{max,t} - SS_{obs,t}) & \text{if } SS_{sim,t} \geq SS_{obs,t} \end{cases} \quad (8)$$

where Q and SS are discharge and suspended sediment load, *sim* is the simulated model result, *obs* is the observed time series and *min* and *max* are the lower and upper LoA uncertainty bounds calculated in section 4.3, respectively, for the time step t. The deviation of the simulated results from the observed series were the smallest when the score was 0 (perfect simulation), ± 1 at the limits of the calculated observational uncertainty and the largest when $\pm\infty$. Therefore, if the simulated prediction is within the calculated LoA of the observation, then the score for that time point will be within the range from 0 to ± 1. If the simulation is beyond the LoA, the score will be greater than ± 1. This resulted in a distribution of scores over the time steps for each model simulation so one can evaluate periods of over- or under-prediction. We use the mean of the absolute scores for all time steps of the selected storm events for S_Q and all time steps during the monitoring period for S_{SS} reported above as our core model simulation performance metric. The top 10% of the best (lowest) mean absolute score were classified as behavioural simulations, and the 5th and 95th percentiles of the behavioural simulations were extracted and presented as the uncertainty bounds of the simulations. These behavioural models were further assigned a conditional probability (*CP*), given the vector of observations (*obs*) as

$$CP\left(R_j(\theta)|Obs\right) = \frac{S_{mean}^{-1} L\left(R_j(\theta)\right)}{\sum_{j=1}^{J} S_{mean}^{-1} L\left(R_j(\theta)\right)} \quad (9)$$

with R_j being one of the $j = 1,\ldots, J$ accepted models with parameter set θ. The conditional probability was used to assess the model parameter identifiability and uncertainty.

To account for the ability of the model to capture the uncertainty limits of observed discharge and cumulative suspended sediment load, two measures were used to evaluate against the resulting 5th and 95th simulation bounds for each time step from the behavioural simulations. The first one is reliability (RM) (Equations 10 and 11), which calculates the overlap between the observed and simulated uncertainty bounds (Westerberg et al., 2011b). RM is calculated as the mean of the percentage of the overlapping range between the observation and simulation relative to the observation and relative to the simulation range, given as

$$RM_Q = \frac{\sum_{t=1}^{T} \left(mean\left(\frac{Q_{Roverlap}}{Q_{Robs}}, \frac{Q_{Roverlap}}{Q_{Rsim}} \right) \right)}{T} \quad (10)$$

$$RM_{SS} = \frac{\sum_{t=1}^{T} \left(mean\left(\frac{SS_{Roverlap}}{SS_{Robs}}, \frac{SS_{Roverlap}}{SS_{Rsim}} \right) \right)}{T} \quad (11)$$

where Q and SS are discharge and suspended sediment load, *Roverlap* is the intersection between the simulated and observed ranges, *Robs* is the observed range and *Rsim* is the simulated range, and T is the number of time steps.

The second one is precision (PM) (Equations 12 and 13), which calculates the average percentage of the width of the overlapping range between the observed and simulated uncertainty bounds to the width of the simulated bounds for all time steps (Guerrero et al., 2013), given as

$$PM_Q = \frac{\sum_{t=1}^{T} \left(\frac{Q_{Roverlap}}{Q_{Rsim}} \right)}{T} \quad (12)$$

$$PM_{SS} = \frac{\sum_{t=1}^{T} \left(\frac{SS_{Roverlap}}{SS_{Rsim}} \right)}{T} \quad (13)$$

The range for both measures is 0–100%. Guidance on a threshold of acceptable values for RM and PM is subjective, so we use these as a diagnostic tool to assess model performance in light of observational uncertainties. However, in general, higher RM values indicate more time on average where the simulated uncertainty bounds overlap with the observational uncertainties, whereas higher PM values mean more time on average where the simulated and observed have similar uncertainty ranges.

5 | RESULTS

5.1 | Model performance to replicate hydrological behaviour

The evaluation on how the behavioural ensemble is able to capture catchment dynamics in reproducing the timing and magnitude of the 14 flood events (10 flood events for Glenderamackin at Threlkeld and St Johns Beck at Thirlmere Reservoir sites) is presented in Figure 5. Note that only the six major flood events, which were representative of the typical range of events, are plotted in the figure for illustration.

Regarding the main gauges of the catchments, the behavioural simulations in Cocker at Southwaite Bridge (Figure 5a) were seen to satisfactorily capture the overall dynamics of the catchment compared with other flow gauging stations, especially for the smaller flood events. However, for most of the major flood events, the ensemble of the behavioural simulations often under-predicted the magnitude of the peaks and did not capture well the recession periods. This was worse for the period of the October 2008 and November 2009 flood events, in which the peaks were under-predicted by about 50% in the Cocker. On the other hand, the behavioural simulations in Derwent at Ouse Bridge (Figure 5b) consistently over-predicted the magnitude of the peaks for the major flood events. The rising limbs of most of the flood events were also consistently over-predicted, with the observed discharge closer to the lower bound of the simulations whereas the recession periods were often the least well captured.

Referring to the gauges within the River Derwent catchment, the ensemble of the behavioural simulations in Derwent at Portinscale (Figure 5c) and in Glendermackin at Threlkeld (Figure 5d) were generally able to capture the timing and magnitude of the peaks, with the observed discharge lying close to the upper bound of the simulations. In the upper part of the catchment in St Johns Beck at Thirlmere

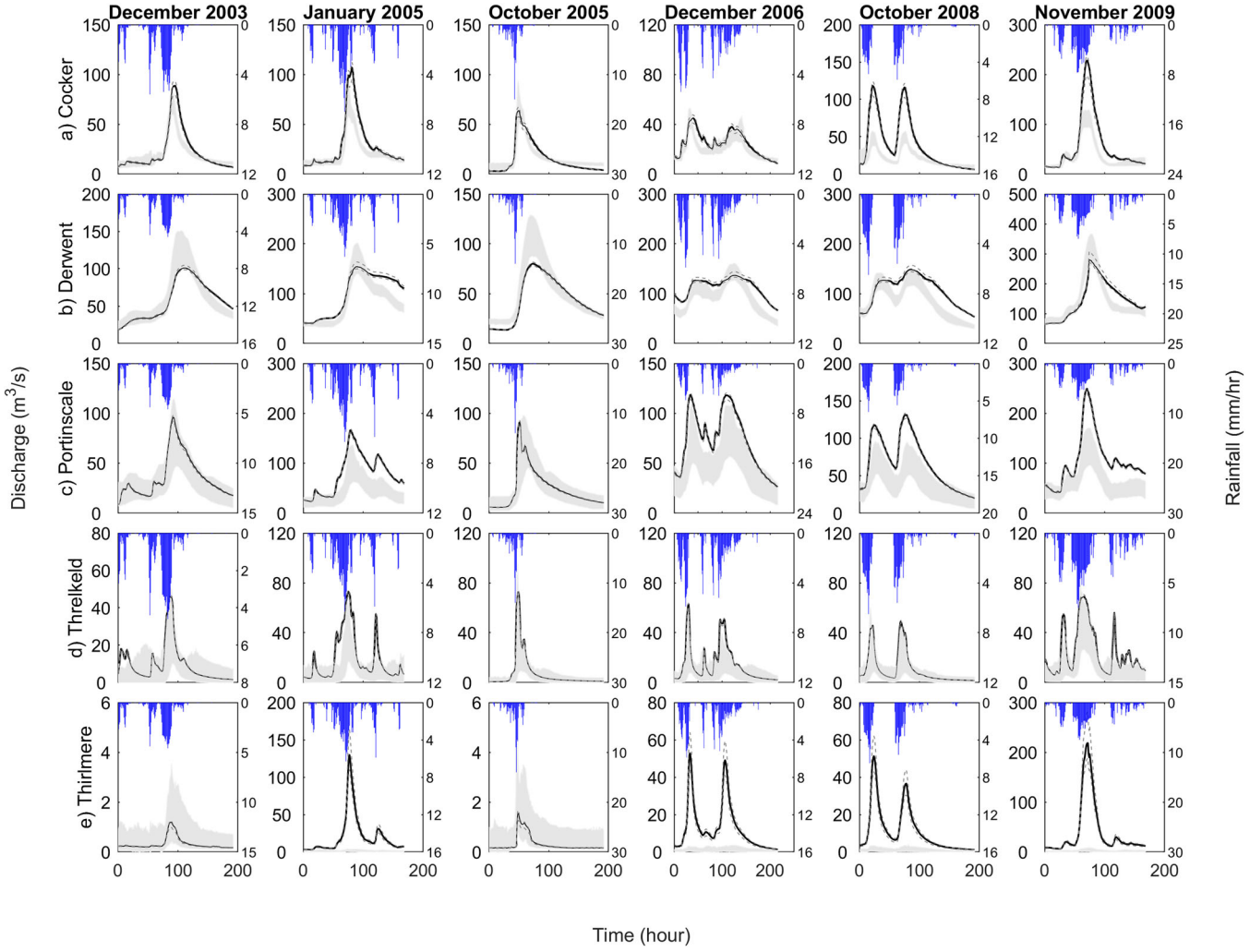

FIGURE 5 Discharge for the 14 flood events at five flow gauging stations. The observed discharge is indicated by black solid line and its uncertainty bounds by grey dashed lines, while the behavioural simulations (5th and 95th percentiles of the simulated discharge) are shown as grey shaded regions. Note that only the six major flood events are plotted in this figure for illustration. The columns indicate the six major flood events, whereas the rows show the five flow gauging stations

Reservoir, the behavioural simulations completely under-predicted the larger events (January 2005, December 2006, October 2008 and November 2009). The major reason why there might be a decline in model performance is the rainfall data being applied uniformly across the whole catchments (see 'Discussion' for further explanation). Another reason might be the artificial regulation of flow from Thirlmere Reservoir because this will particularly affect the magnitude of flood peaks, which is dependent on the reservoir level. The flow measured at the station 1 km downstream of the reservoir is affected by public water supply abstraction and also flood release regulation.

5.2 | Model performance to reproduce sediment yields

Figure 6 illustrates how the behavioural ensemble is able to capture the sediment dynamics at the two monitoring sites. In addition, the cumulative suspended sediment loads were plotted (Figure 7) to examine whether the behavioural models were able to capture the timing and magnitude of the whole time series in terms of accumulation.

At both monitoring sites, the majority of the sediment loads at the low-load evaluation points were totally under-predicted at or greater than 10% exceedance (Figure 6, left panel) and at or greater than 1% exceedance (Figure 6, right panel) for the Derwent at Portinscale and for Newlands Beck at Braithwaite, respectively. This shows that the behavioural simulations did not generate any suspended sediment loads for ≥95% of the time but were able to capture the high-load part of the sediment load–duration curve during effective discharge events.

In Figure 7, the cumulative suspended sediment load in Derwent at Portinscale lies within the wider range of the 5th and 95th percentiles of the simulations before November, whereas the behavioural simulations in Newlands Beck at Braithwaite totally under-predicted the cumulative suspended sediment load, especially during the individual events before August. This reveals that the behavioural simulations in Derwent at Portinscale were highly variable and inconsistent among each other because the high-load events were produced at different times during the monitoring period. Similar sediment yields simulated at the gauge location could be results of very different behaviours within the catchment. On the other hand, the behavioural simulations in Newlands Beck at Braithwaite were more consistent

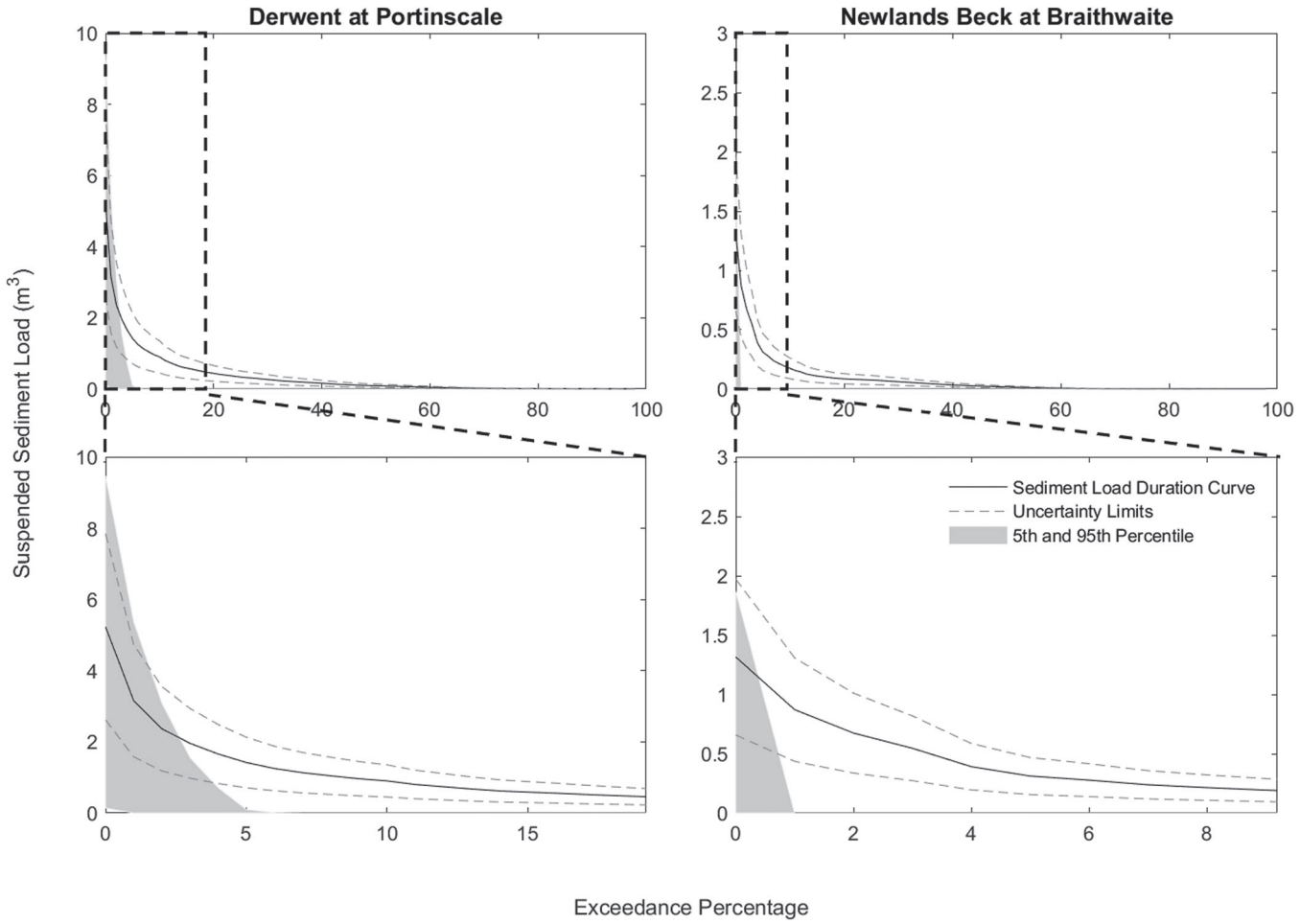

FIGURE 6 Sediment load–duration curve for the two monitoring sites in Derwent at Portinscale (left) and in Newlands Beck at Braithwaite (right). The observed sediment load–duration curve is indicated by black solid line and its uncertainty limits by grey dashed lines, whereas the behavioural simulations (5th and 95th percentiles of the simulated suspended sediment load) are shown as grey shaded regions. Top panel: the whole range of exceedance percentage; bottom panel: zoom-in image at or greater than 20% (left) and at or greater than 10% (right) exceedance percentages

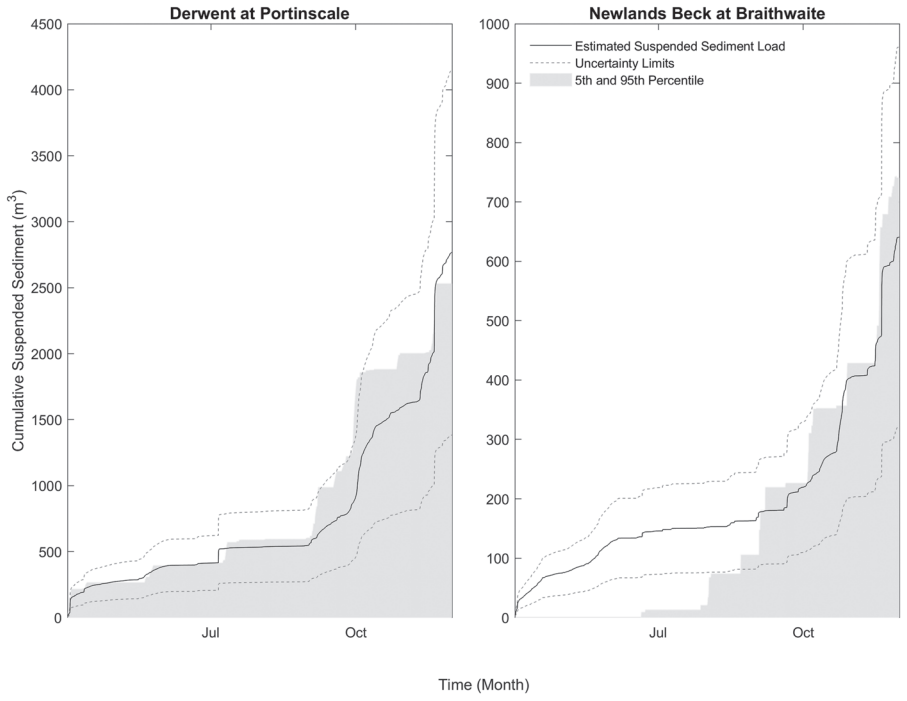

FIGURE 7 Cumulative suspended sediment load (m^3) for the two monitoring sites in Derwent at Portinscale (left) and in Newlands Beck at Braithwaite (right) during the period of 1 April to 30 November 2006. The observed cumulative suspended sediment load is indicated by black solid line and its uncertainty bounds by grey dashed lines, whereas the behavioural simulations (5th and 95th percentiles of the simulated cumulative suspended sediment load) are shown as grey shaded regions

than those in Derwent at Portinscale, generating the suspended sediment loads mainly after September. In general, the accumulation of the suspended sediment loads produced by the behavioural simulations at both monitoring sites was under-estimated by the end of the monitoring period.

5.3 | Behavioural ensembles for different diagnostics

Based on the behavioural simulations, the reliability (RM) and precision measures (PM) for the hydrological response were calculated and are summarized in Table 4. These reveal the overall ability of the behavioural simulations to capture the uncertainty limits of observed discharge for all 14 flood events (10 flood events for Glenderamackin at Threlkeld and St Johns Beck at Thirlmere Reservoir).

For the Cocker catchment, the simulations at this gauging station performed consistently for RM and PM because the three highest RM (51.1%, 42.7% and 52.2%) were associated with the three highest PM (20.2%, 18.4% and 17.6%) in 2001, 2006 and 2011, respectively. This was also the case with the three lowest RMs (31.5%, 24.0% and 19.6%) being associated with the three lowest PMs (10.5%, 10.8% and 7.7%) in 2002, October 2005 and 2008, respectively. For the River Derwent catchment, the RM and PM showed an interesting pattern in terms of spatial variability. The overall RM increased generally from the lower catchment (Derwent at Ouse Bridge: 25.0%) to the upper catchment (Glenderamackin at Threlkeld: 41.2%), with a decreasing average PM (from 8.4% to 2.6%). Although St Johns Beck at Thirlmere Reservoir is located at the upper part of the River Derwent catchment, the simulations at this gauging station had the poorest performance: the overall lowest RM (24.4%) and the second lowest average PM (3.7%). Such spatial variability could be attributed to the fact that a wider range of simulation bounds provided higher chance of overlapping with the observation uncertainty limits with less precision.

The overall ability of the behavioural simulations to replicate the uncertainty bounds of the sediment load–duration curve is also summarized by the RM and PM. The 5th and 95th behavioural simulations in Derwent at Portinscale were seen to better match with the uncertainty limits of the observations compared with those produced in Newlands Beck at Braithwaite, as shown by both higher RM and PM of the former (2.8%; 12.7%) than the latter (0.8%; 6.5%).

In general, the behavioural simulations were better in overlapping with the uncertainty limits of the streamflow than those of the suspended sediment loads. However, care must be taken because the time step-based performance measures could be affected by the sampling errors. Due to the heavy computational demand of the model, only 1,500 simulations were run, and these simulations did not represent a dense sample when sampling a model space of 15 parameters. The measures were also susceptible to the effects of observation errors. It was, therefore, not surprising that all model simulations did not fall within the observed uncertainty limits for all time steps.

5.4 | Assessing model parameter uncertainty and equifinality

Figures 8 and 9 show the dotty plots for each sampled parameter for the behavioural simulations based on the conditional probability in the River Derwent and Cocker catchments, respectively. Three of the

TABLE 4 RM and PM for hydrological metrics in each flood event at five flow gauging stations

Gauge number		75,004		75,003		75,005		75,007		75,001	
River		Cocker		Derwent		Derwent		Glenderamackin		St Johns Beck	
Gauge station		Southwaite bridge		Ouse bridge		Portinscale		Threlkeld		Thirlmere reservoir	
Measure (%)		RM	PM	RM	PM	RM	PM	RM	PM	RM	PM
Flood event	1999	35.1	13.7	10.7	2.6	12.3	1.7	--	--	--	--
	2000	38.8	14.4	34.8	*12.2*	38.1	5.0	--	--	--	--
	2001	*51.1*	*20.2*	25.1	8.8	*44.1*	*5.9*	--	--	--	--
	2002	31.5	10.5	*56.1*	*14.7*	30.5	4.4	--	--	--	--
	2003*	41.2	13.6	*35.4*	6.3	*44.0*	*6.0*	*45.7*	*2.1*	50.3	0.7
	2004	32.5	14.1	19.7	5.2	30.0	3.7	45.7	*2.2*	45.0	9.4
	Jan 2005*	39.1	13.4	28.9	8.6	9.5	1.4	38.4	2.1	19.6	8.1
	Oct 2005*	24.0	10.8	28.1	8.7	*51.7*	*7.0*	*48.8*	1.5	17.4	0.4
	2006*	*42.7*	*18.4*	22.5	8.1	9.8	1.5	40.6	2.1	45.6	*10.9*
	2007	32.3	13.1	*42.7*	*10.3*	29.3	4.0	*45.8*	1.5	*51.8*	*11.2*
	2008*	19.6	7.7	22.3	7.2	11.1	1.6	44.6	1.2	36.5	*11.2*
	2009*	33.5	12.8	29.4	9.2	12.4	1.9	36.8	*2.7*	12.3	5.5
	2010	41.4	10.9	33.4	7.5	40.3	5.5	43.8	1.2	*52.8*	5.6
	2011	*52.2*	*17.6*	27.5	5.7	41.2	5.5	44.0	1.0	*50.7*	1.4
	Overall	**33.0**	**16.2**	**25.0**	**8.4**	**34.6**	**5.0**	**41.2**	**2.6**	**24.4**	**3.7**

*Six major flood events that happened in the catchment during the observation period. The three highest RM and PM in each flow gauging station are indicated in *italic*, whereas the three lowest are underlined

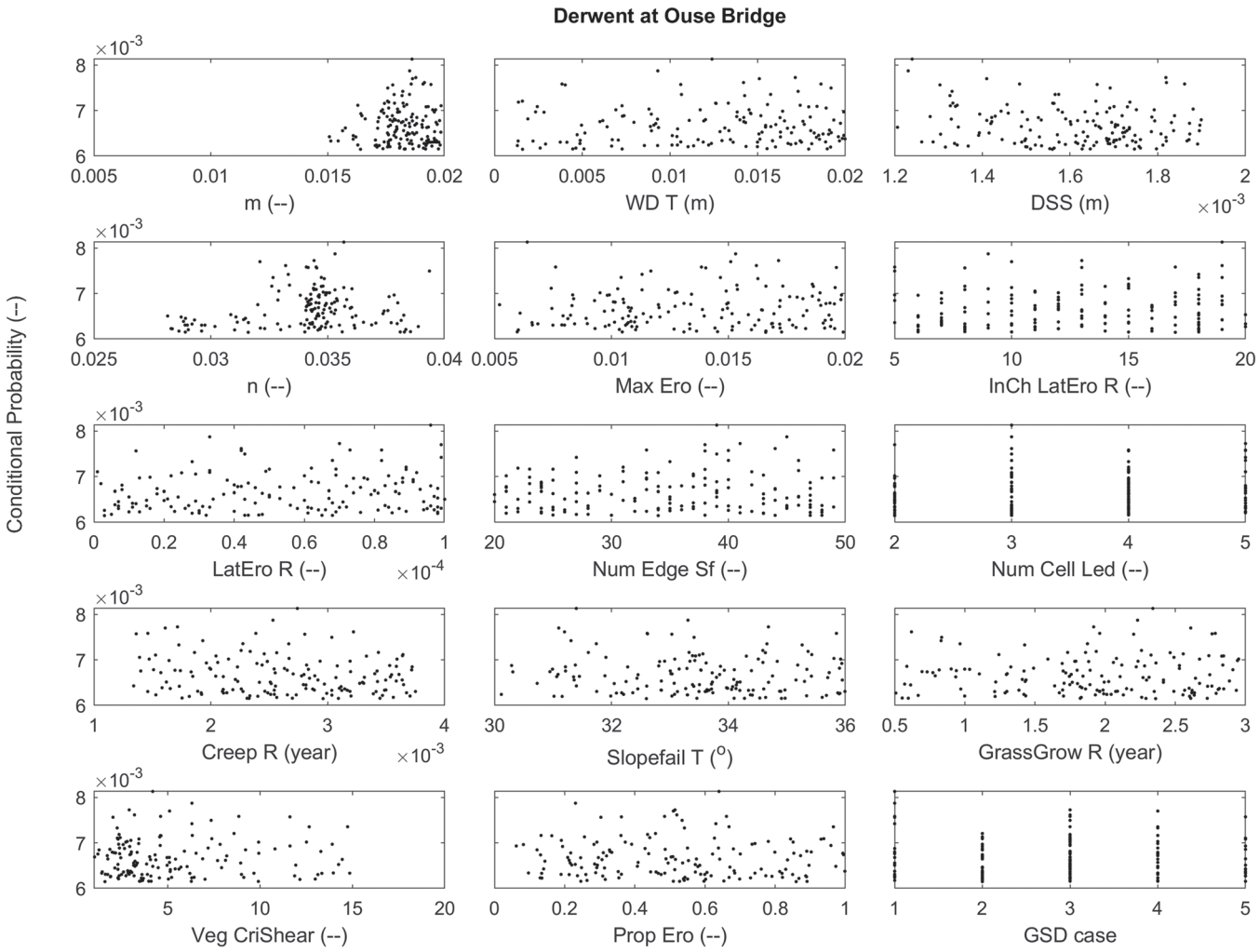

FIGURE 8 Dotty plots of behavioural parameter sets for the 15 CAESAR-LISFLOOD parameters (the parameter names are explained in Table 3) in the River Derwent catchment. The conditional probability was calculated using equation 9

parameters (*m*, *n* and *Veg CriShear*) have been shown to be more identifiable relative to the original range sampled. The hydrologic parameter *m* was the most identifiable parameter for both catchments. This parameter is an important variable in catchment-mode CAESAR-Lisflood as it controls the peak and duration of the hydrograph generated by a rainfall event. A low value means higher and flashier peaks, and vice versa. More favourable results were found in higher values (0.018–0.019) of the parameter space in River Derwent and lower values (0.008–0.009) in Cocker. This could be explained by the fact that the River Cocker has a more flashy response than the River Derwent and the effect of Bassenthwaite Lake, in which the flashy response of the upstream reaches is attenuated by the lake in the downstream sections of the River Derwent. The Manning friction coefficient *n* was shown to have favourable results found in a range of values between 0.031 to 0.036 in both catchments, probably because of the friction coefficient being a lumped value for both channel and floodplain friction. Also, the vegetation parameter *Veg_CriShear* was identified to have values less than 16, with a tendency for more favourable results found with values less than 5. The behavioural simulations of the remaining parameters were distributed across the original parameter ranges, thus they were considered as non-identifiable parameters. This suggests the possible presence of equifinality, which could originate from parameter correlations and interactions, imperfect knowledge of the system under consideration, and different sources of error (e.g. input errors, model structural errors and observational errors) that interacted in a non-linear way.

6 | DISCUSSION

The purpose of this modelling exercise is to examine the capability of CAESAR-Lisflood in capturing the hydrological and geomorphic dynamics of the River Derwent and Cocker catchments during floods, with the consideration of different observational uncertainties within the stage–discharge and suspended sediment load data. Given the different nature and level of uncertainties in the observed data, a decline in model performance was shown from the lower to upper part of the catchment. This could be mainly because the rainfall data were applied uniformly across the whole study area, even though some improvements were made by calculating the areal average rainfall of the catchment and by reproducing the stochastic nature of the rainfall series for the spin-up period. The homogeneity of spatial rainfall distribution in the model could have significant effects on runoff generation, especially in the upper part of the catchment where orographic effects occur. It could be expected that the locally intense storm events and subsequent runoff events would be under-represented by

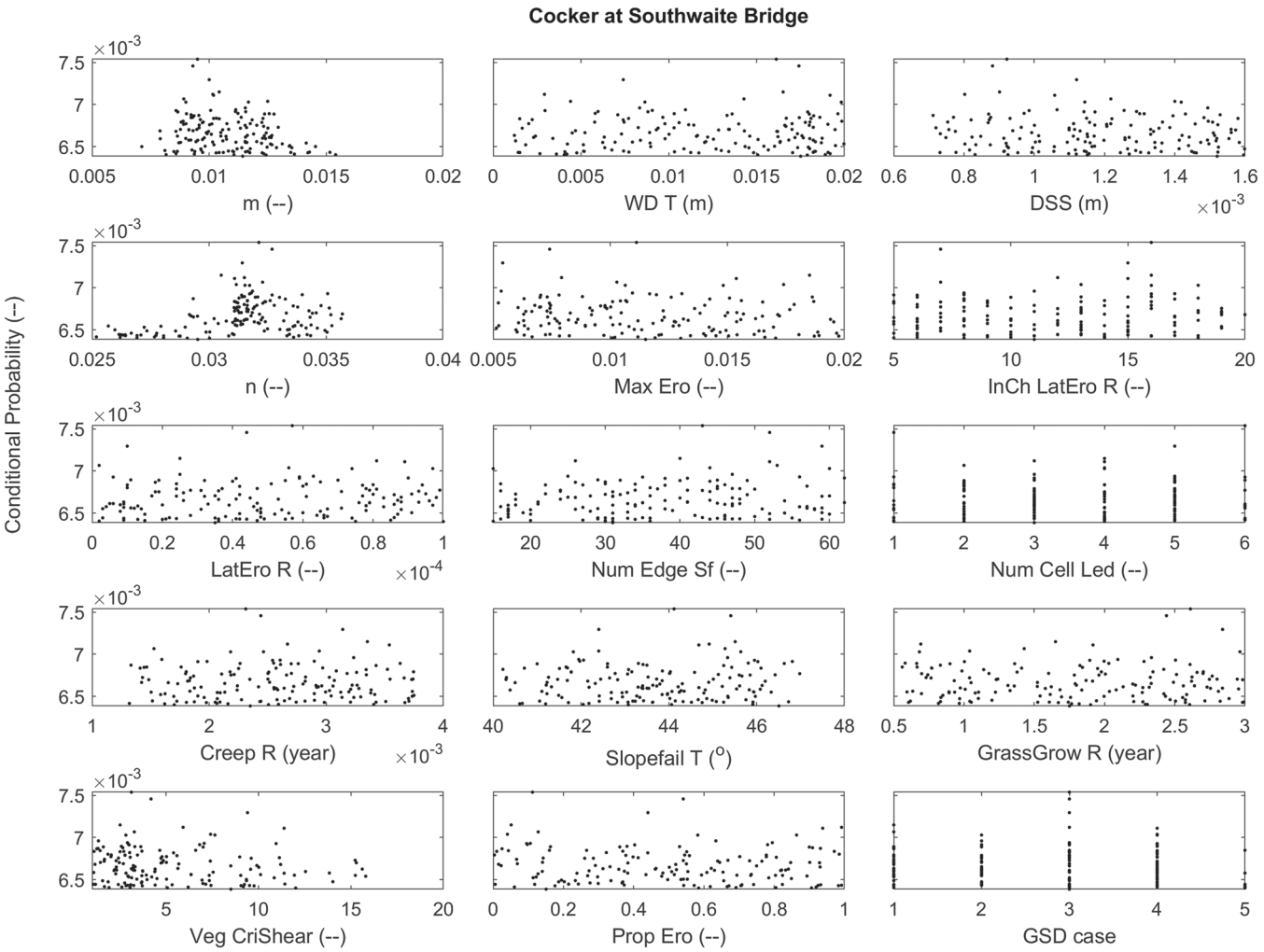

FIGURE 9 Dotty plots of behavioural parameters for the 15 CAESAR-LISFLOOD parameters (the parameter names are explained in Table 3) in the Cocker catchment. The conditional probability was calculated using equation 9

the homogeneous rainfall. The spatial rainfall aspects of the model were only evaluated recently, and their impact remains uncertain. Coulthard and Skinner (2016) demonstrated that the temporal and spatial resolutions of the rainfall data had a small impact on basin hydrology but generated larger changes in basin geomorphology. Though subsequent simulations showed CAESAR-Lisflood simulations were sensitive to the choice of rainfall products, where hydrological changes were linear but non-linear differences were seen in sediment yields (Skinner et al., 2020). Therefore, the impacts of spatial rainfall input on basin discharge in larger basins like ours remains as a hypothesis to be tested. More in-depth analyses could be done to further investigate the impacts of rainfall input on model performance, but this is not within the scope of this study and future work could be done, for instance, to fully account for the rainfall uncertainties arising from the measurement errors and the spatial representation of the rain gauges.

The model performance in representing the hydrological and geomorphic behaviours of the catchments could also be affected by missing/simplification of process representation. As both catchments have several lakes, the lakes were treated as large depressions in the DEM, where sediments were deposited at the entrance and piled up further in the lower part of the depressions over time. Water was routed through these depressions, which filled up until the flow found its way to the exit. The lack of lake dynamic processes in the model would particularly affect the simulations in Derwent at Ouse Bridge (downstream of Bassenthwaite Lake) and St Johns Beck at Thirlmere Reservoir. Similarly, bank erosion was identified as the significant component of the overall observed sediment load, of which part of the loads was supplemented by temporary in-channel storage of fine sediment that was re-entrained during high flow (Warburton, 2010). The simplified representation of lateral erosion processes (four parameters related to bank failure) in the model and the coarse cell resolution (200 m in the River Derwent and 100 m in the River Cocker) in which sub-grid scale processes could not be fully represented (e.g. the exact characteristics of river channels) could have impacts on the ability of the model in capturing smaller local events (e.g. bank failure). A comparison of the geomorphic statistics (area–slope relationship and hypsometric curve and integral) showed that there were limited differences among the DEMs with different resolutions, especially given the uncertainty spread in some of these relationships (Figures A1 and A2). However, the behavioural simulations were still able to capture the high-magnitude low-frequency sediment events, which were possibly in terms of effective geographic events such as landslides or hydrological events such as flooding (Warburton et al., 2008). With the current focus on the model ability in simulating hydrological and geomorphic dynamics of the catchments during

floods, we argue that the model was able to provide realistic simulations and their associated uncertainties for subsequent analysis (e.g. as inputs for downstream flood analysis), given the model performance at catchment outlets showing better behaviours.

It is acknowledged that different sets of behavioural simulations were found when evaluating the model performance with different gauges within the catchment, different flood events and different types of observed data. There was no single behavioural simulation that could adequately and simultaneously reproduce both hydrological and sedimentological behaviours across different parts of the catchment. In many ways, this is not surprising because of the highly non-linear relationship between rainfall and discharge/sediment yield. This non-linear hydrology–sediment delivery response is fundamentally embedded into the sediment transport formulae in the model, such that the shear stress is proportional to the square of the flow velocity, and the flow velocity is non-linearly related to discharge as well (Coulthard & Van De Wiel, 2007). Given the uncertainties within the suspended sediment measurements and the stage–discharge rating curves, it remains a challenge to evaluate LEM performance in quantifying and assessing both at-a-point behaviour (e.g. sediment yield) and areal catchment response (e.g. streamflow). We believe that addressing these uncertainties directly in model performance metrics is necessary, particularly as each type of data has its own error characteristics. Instead of assuming a good degree of data quality and using standard objective functions, we should be more explicit about quantifying data uncertainties when evaluating the model performance. While there are increasing numbers of research papers in the hydrological modelling literature considering observational uncertainties in model assessments, we advocate that similar awareness is needed in assessing the LEMs, and our study could be regarded as the beginning of a dialogue to deal with data uncertainties in LEM evaluation.

The LoA approach adopted in this study provided a flexible way to explore the interactions between observational uncertainty and parameter uncertainty. This is an important step looking at the uncertainty of the model performance rather than the sensitivity. Given the different nature and level of uncertainties in the observed data, the hydrologic parameter (*m*), Manning friction coefficient (*n*) and vegetation critical shear stress (*Veg CriShear*) were shown to be more identifiable for both catchments in this study, which aligned with the findings concluded by previous studies (Skinner et al., 2018; Welsh et al., 2009). Similar model performance by different behavioural parameter sets revealed that simulated results at a specific location could be driven by similar or totally different dynamics within the catchment (i.e. equifinality due to parameter uncertainties). This also implies that simulated results at a specific location of the catchment (e.g. catchment outlet) could only reflect the catchment behaviour of that specific location but do not necessarily provide the relevant information on other parts of the catchment (Skinner et al., 2018). Even with a large amount of available multi-proxy data, the reliability of the LEM was evaluated with the assumption of a good degree of data quality and low impact of other error sources (Gioia & Lazzari, 2019). In this regard, our study applied for the first time the LoA approach within a GLUE uncertainty analysis framework to assess the reliability of a LEM with the consideration of observational and parameter uncertainties. This highlights a high potential of using such an approach in assessing different sources of uncertainty in LEMs.

7 | CONCLUSIONS

Despite their proven capabilities, the success of LEMs has been hampered by the lack of uncertainty investigations, in most cases because of the paucity of available data. Even when detailed evaluation data are available, the uncertainties inherent in observed data (e.g. discharge and suspended sediment load) could be largely owing to the unpredictability of the environmental time series and errors in collection techniques. Thus, a framework of the LoA approach in GLUE was adopted in this study, which enables parameter conditioning under the circumstances of highly uncertain data. This also allows for integration of different types of data in the parameter conditioning process by reflecting their quality. Here, we assessed the ability of hydrological and geomorphic uncertain observations in constraining catchment response and to explore the parameter identifiability in a landscape evolution model, CAESAR-Lisflood, using the River Derwent and Cocker catchments in the UK as an example.

The evaluation of the model performance in different locations of River Derwent catchment provides an opportunity to reveal the spatially heterogeneous responses of different flood events. Given the limitations in the rainfall inputs, simplified process representation and coarse resolution of the model, the resultant 5th and 95th simulation bounds of the behavioural models are still able to provide considerable overlap with the observation uncertainty limits (24.4–41.2% of reliability measures). However, the ability of the model simulations to reproduce observations is dependent on which constraints are applied to the model by the user and how well the model users define the effective observation errors. It is more likely that the rainfall uncertainty and the spatial representation of rainfall fields could have a larger impact on the model performance and on acceptable effective parameter values (i.e. Skinner et al., 2020).

Given the incomplete understanding of the data uncertainties and inherent limitations in the model setting, future work should be undertaken to focus on quantifying the rainfall input uncertainty as part of the model evaluation and examining the effects of spatially distributed rainfall input on the capability of model to produce realistic simulations and catchment behaviour. This is important to investigate how all these uncertainties in upstream propagate to downstream when the results are subsequently used for later analysis of flood predictions at Cockermouth, where the channel undergoes geomorphic change during large floods (Wong et al., 2015). Finally, although the sophistication of landscape evolution models is rapidly evolving, evaluating their performance is often limited by the availability of long-duration, high-resolution field datasets and spatial data. This paper is a first step to full-ensemble uncertainty evaluation of such models that reflects the challenges of constraining simulations with uncertain observations.

ACKNOWLEDGEMENTS

Jefferson Wong was funded by a University of Bristol Postgraduate Scholarship. The authors thank the UK Environment Agency for providing the rain gauge and flow gauging station data. Thanks are also due to Dr. Jorge Ramirez for his efforts in collecting the grain size data and providing access to the data used in this study. The authors furthermore thank the two anonymous reviewers and Stuart Lane, the editor, for their invaluable comments, which greatly helped to improve the quality of the paper. Paul Bates was supported by a Royal Society

Wolfson Research Merit Award. Jim Freer was partly funded for his time by the Global Water Futures program, University of Saskatchewan. Paul Bates and Jim Freer time was partly funded by NERC grant NE/K00882X/1.

CONFLICT OF INTEREST

There is no conflict of interest in this paper.

DATA AVAILABILITY STATEMENT

Data used in this study are available from the corresponding author upon reasonable request.

ORCID

Jefferson S. Wong https://orcid.org/0000-0002-6793-5017
Tom J. Coulthard https://orcid.org/0000-0002-8164-0442

REFERENCES

Ajami, N.K., Duan, Q.Y. & Sorooshian, S. (2007) An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43(1), W01403. https://doi.org/10.1029/2005wr004745

Armitage, J.J., Whittaker, A.C., Zakari, M. & Campforts, B. (2018) Numerical modelling of landscape and sediment flux response to precipitation rate change. *Earth Surface Dynamics*, 6(1), 77–99. https://doi.org/10.5194/esurf-6-77-2018

Asselman, N.E.M. (2000) Fitting and interpretation of sediment rating curves. *Journal of Hydrology*, 234(3–4), 228–248. https://doi.org/10.1016/S0022-1694(00)00253-5

Bastola, S., Dialynas, Y.G., Bras, R.L., Noto, L.V. & Istanbulluoglu, E. (2018) The role of vegetation on gully erosion stabilization at a severely degraded landscape: A case study from Calhoun Experimental Critical Zone Observatory. *Geomorphology*, 308, 25–39. https://doi.org/10.1016/j.geomorph.2017.12.032

Bates, P.D., Horritt, M.S. & Fewtrell, T.J. (2010) A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, 387(1–2), 33–45. https://doi.org/10.1016/j.jhydrol.2010.03.027

Beven, K. (1996) Equifinality and uncertainty in geomorphological modelling. *Scientific Nature of Geomorphology*, 27, 289–313.

Beven, K. (2006) A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. & Binley, A. (1992) The future of distributed models - model calibration and uncertainty prediction. *Hydrological Processes*, 6(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K. & Freer, J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K.J. (2009) *Environmental modelling: an uncertain future?: An introduction to techniques for uncertainty estimation in environmental prediction*. London; New; York: Routledge.

Beven, K.J. & Kirkby, M.J. (1979) A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, 24(1), 43–69. https://doi.org/10.1080/02626667909491834

Blazkova, S. & Beven, K. (2009) A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, 45(12), W00b16. https://doi.org/10.1029/2007wr006726

Cameron, D.S., Beven, K.J., Tawn, J., Blazkova, S. & Naden, P. (1999) Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). *Journal of Hydrology*, 219(3–4), 169–187. https://doi.org/10.1016/S0022-1694(99)00057-8

Chiverrell, R.C., Sear, D.A., Warburton, J., Macdonald, N., Schillereff, D.N., Dearing, J.A. *et al.* (2019) Using lake sediment archives to improve understanding of flood magnitude and frequency: Recent extreme flooding in northwest UK. *Earth Surface Processes and Landforms*, 44(12), 2366–2376. https://doi.org/10.1002/esp.4650

Chow, V.T. (1959) *Open Channel Hydraulics*. New York: McGraw-Hill, pp. 26–27.

Claessens, L., Heuvelink, G.B.M., Schoorl, J.M. & Veldkamp, A. (2005) DEM resolution effects on shallow landslide hazard and soil redistribution modelling. *Earth Surface Processes and Landforms*, 30(4), 461–477. https://doi.org/10.1002/esp.1155

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829. https://doi.org/10.2307/2286407

Collins, D.B.G., Bras, R.L. & Tucker, G.E. (2004) Modeling the effects of vegetation-erosion coupling on landscape evolution. *Journal of Geophysical Research - Earth Surface*, 109, F03004. https://doi.org/10.1029/2003jf000028

Coulthard, T.J., Kirkby, M.J. & Macklin, M.G. (2000) Modelling geomorphic response to environmental change in an upland catchment. *Hydrological Processes*, 14(11–12), 2031–2045. https://doi.org/10.1002/1099-1085(20000815/30)14:11/12%3C2031::Aid-Hyp53%3E3.0.Co;2-G

Coulthard, T.J., Lewin, J. & Macklin, M.G. (2005) Modelling differential catchment response to environmental change. *Geomorphology*, 69(1–4), 222–241. https://doi.org/10.1016/j.geomorph.2005.01.008

Coulthard, T.J. & Macklin, M.G. (2001) How sensitive are river systems to climate and land-use changes? A model-based evaluation. *Journal of Quaternary Science*, 16(4), 347–351. https://doi.org/10.1002/jqs.604

Coulthard, T.J. & Macklin, M.G. (2003) Modeling long-term contamination in river systems from historical metal mining. *Geology*, 31(5), 451. https://doi.org/10.1130/0091-7613(2003)031%3C0451:Mlcirs%3E2.0.Co;2

Coulthard, T.J., Macklin, M.G. & Kirkby, M.J. (2002) A cellular model of Holocene upland river basin and alluvial fan evolution. *Earth Surface Processes and Landforms*, 27(3), 269–288. https://doi.org/10.1002/esp.318

Coulthard, T.J., Neal, J.C., Bates, P.D., Ramirez, J., de Almeida, G.A.M. & Hancock, G.R. (2013) Integrating the LISFLOOD-FP 2D hydrodynamic model with the CAESAR model: implications for modelling landscape evolution. *Earth Surface Processes and Landforms*, 38(15), 1897–1906. https://doi.org/10.1002/esp.3478

Coulthard, T.J. & Skinner, C.J. (2016) The sensitivity of landscape evolution models to spatial and temporal rainfall resolution. *Earth Surface Dynamics*, 4(3), 757–771. https://doi.org/10.5194/esurf-4-757-2016

Coulthard, T.J. & Van De Wiel, M.J. (2007) Quantifying fluvial non linearity and finding self organized criticality? Insights from simulations of river basin evolution. *Geomorphology*, 91(3–4), 216–235. https://doi.org/10.1016/j.geomorph.2007.04.011

Coxon, G., Freer, J., Westerberg, I.K., Wagener, T., Woods, R. & Smith, P.J. (2015) A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, 51(7), 5531–5546. https://doi.org/10.1002/2014wr016532

Densmore, A.L., Ellis, M.A. & Anderson, R.S. (1998) Landsliding and the evolution of normal-fault-bounded mountains. *Journal of Geophysical Research - Solid Earth*, 103(B7), 15203–15219. https://doi.org/10.1029/98jb00510

Duan, Q.Y., Ajami, N.K., Gao, X.G. & Sorooshian, S. (2007) Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, 30(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Efron, B. (1979) 1977 Rietz lecture - bootstrap methods - another look at the jackknife. *Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Environment Agency. (2006). Table of lake facts. Access from http://webarchive.nationalarchives.gov.uk/20060806052432/http://www.environment-agency.gov.uk/commondata/103196/1118608?

referrer=/regions/northwest/346910/347005/1110630/1110650/ on 11th Oct, 2013.

Environment Agency. (2009). River Derwent Catchment Flood Management Plan: Summary Report. Environment Agency.

Ferguson, R.I. (1987) Accuracy and precision of methods for estimating river loads. *Earth Surface Processes and Landforms*, 12(1), 95–104. https://doi.org/10.1002/esp.3290120111

Folk, R.L. & Ward, W.C. (1957) Brazos River bar [Texas]; a study in the significance of grain size parameters. *Journal of Sedimentary Research*, 27(1), 3–26. https://doi.org/10.1306/74D70646-2B21-11D7-8648000102C1865D

Gioia, D. & Lazzari, M. (2019) Testing the prediction ability of LEM-derived sedimentary budget in an upland catchment of the Southern Apennines, Italy: a source to sink approach. *Water*, 11(5), 911. https://doi.org/10.3390/w11050911

Goovaerts, P. (2000) Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1–2), 113–129. https://doi.org/10.1016/S0022-1694(00)00144-X

Guerrero, J.L., Westerberg, I.K., Halldin, S., Lundin, L.C. & Xu, C.Y. (2013) Exploring the hydrological robustness of model-parameter values with alpha shapes. *Water Resources Research*, 49(10), 6700–6715. https://doi.org/10.1002/wrcr.20533

Hancock, G.R. (2009) A catchment scale assessment of increased rainfall and storm intensity on erosion and sediment transport for Northern Australia. *Geoderma*, 152(3–4), 350–360. https://doi.org/10.1016/j.geoderma.2009.07.003

Hancock, G.R., Coulthard, T.J. & Lowry, J.B.C. (2016) Predicting uncertainty in sediment transport and landscape evolution - the influence of initial surface conditions. *Computers & Geosciences*, 90, 117–130. https://doi.org/10.1016/j.cageo.2015.08.014

Hancock, G.R., Evans, K.G., Willgoose, G.R., Moliere, D.R., Saynor, M.J. & Loch, R.J. (2000) Medium-term erosion simulation of an abandoned mine site using the SIBERIA landscape evolution model. *Australian Journal of Soil Research*, 38(2), 249. https://doi.org/10.1071/Sr99035

Hancock, G.R., Lowry, J.B.C. & Coulthard, T.J. (2016) Long-term landscape trajectory - Can we make predictions about landscape form and function for post-mining landforms? *Geomorphology*, 266, 121–132. https://doi.org/10.1016/j.geomorph.2016.05.014

Hancock, G.R., Verdon-Kidd, D. & Lowry, J.B.C. (2017) Soil erosion predictions from a landscape evolution model - An assessment of a post-mining landform using spatial climate change analogues. *Science of the Total Environment*, 601, 109–121. https://doi.org/10.1016/j.scitotenv.2017.04.038

Hancock, G.R., Willgoose, G.R. & Evans, K.G. (2002) Testing of the SIBERIA landscape evolution model using the Tin Camp Creek, Northern Territory, Australia, field catchment. *Earth Surface Processes and Landforms*, 27(2), 125–143. https://doi.org/10.1002/esp.304

Hatfield, R.G. & Maher, B.A. (2008) Suspended sediment characterization and tracing using a magnetic fingerprinting technique: Bassenthwaite Lake, Cumbria, UK. *Holocene*, 18(1), 105–115. https://doi.org/10.1177/0959683607085600

Hatfield, R.G. & Maher, B.A. (2009) Holocene sediment dynamics in an upland temperate lake catchment: climatic and land-use impacts in the English Lake District. *Holocene*, 19(3), 427–438. https://doi.org/10.1177/0959683608101392

Hicks, D.M., Gomez, B. & Trustrum, N.A. (2000) Erosion thresholds and suspended sediment yields, Waipaoa River Basin, New Zealand. *Water Resources Research*, 36(4), 1129–1142. https://doi.org/10.1029/1999wr900340

Hollaway, M.J., Beven, K.J., Benskin, C.M.H., Collins, A.L., Evans, R., Falloon, P.D. *et al.* (2018) The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model. *Journal of Hydrology*, 558, 607–624. https://doi.org/10.1016/j.jhydrol.2018.01.063

Hooke, J.M., Brookes, C.J., Duane, W. & Mant, J.M. (2005) A simulation model of morphological, vegetation and sediment changes in ephemeral streams. *Earth Surface Processes and Landforms*, 30(7), 845–866. https://doi.org/10.1002/esp.1195

Ijjaszvasquez, E.J., Bras, R.L. & Moglen, G.E. (1992) Sensitivity of a basin evolution model to the nature of runoff production and to initial conditions. *Water Resources Research*, 28(10), 2733–2741. https://doi.org/10.1029/92wr01561

Kiang, J.E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I.K. *et al.* (2018) A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, 54(10), 7149–7176. https://doi.org/10.1029/2018wr022708

Krueger, T., Freer, J., Quinton, J.N., Macleod, C.J.A., Bilotta, G.S., Brazier, R.E. *et al.* (2010) Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, 46(7), W07516. https://doi.org/10.1029/2009wr007845

Krzysztofowicz, R. (2002) Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology*, 268(1–4), 16–40. https://doi.org/10.1016/S0022-1694(02)00106-3

Kuczera, G. & Parent, E. (1998) Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology*, 211(1–4), 69–85. https://doi.org/10.1016/S0022-1694(98)00198-X

Kwang, J.S. & Parker, G. (2019) Extreme memory of initial conditions in numerical landscape evolution models. *Geophysical Research Letters*, 46(12), 6563–6573. https://doi.org/10.1029/2019gl083305

Liu, Y.L., Freer, J., Beven, K. & Matgen, P. (2009) Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, 367(1–2), 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016

Mackay, J.D., Barrand, N.E., Hannah, D.M., Krause, S., Jackson, C.R., Everest, J. & Aoalgeirsdottir, G. (2018) Glacio-hydrological melt and run-off modelling: application of a limits of acceptability framework for model comparison and selection. *The Cryosphere*, 12(7), 2175–2210. https://doi.org/10.5194/tc-12-2175-2018

Mair, A. & Fares, A. (2011) Comparison of rainfall interpolation methods in a mountainous region of a tropical island. *Journal of Hydrologic Engineering*, 16(4), 371–383. https://doi.org/10.1061/(Asce)He.1943-5584.0000330

Mantovan, P. & Todini, E. (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology*, 330(1–2), 368–381. https://doi.org/10.1016/j.jhydrol.2006.04.046

McMillan, H., Freer, J., Pappenberger, F., Krueger, T. & Clark, M. (2010) Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, 24, 1270–1284. https://doi.org/10.1002/hyp.7587

McMillan, H., Krueger, T. & Freer, J. (2012) Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, 26(26), 4078–4111. https://doi.org/10.1002/hyp.9384

Mill, H.R. (1895) Bathymetrical survey of the English Lakes. *The Geographical Journal*, 6(1), 46–73. https://doi.org/10.2307/1773952

Montanari, A. (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, 41(8), W08406. https://doi.org/10.1029/2004wr003826

Moseley, F. (1979). The Geology of the Lake District: Yorkshire Geological Society Occasional Publication, no. 3.

Moyeed, R.A. & Clarke, R.T. (2005) The use of Bayesian methods for fitting rating curves, with case studies. *Advances in Water Resources*, 28(8), 807–818. https://doi.org/10.1016/j.advwatres.2005.02.005

Owens, P. & Collins, A. (2006) Soil Erosion and Sediment Redistribution in River Catchments Measurement, Modelling and Management Preface. In: *Soil Erosion and Sediment Redistribution in River Catchments: Measurement, Modelling and Management*. Wallingford, Oxfordshire, UK: CABI, pp. Xii–Xiv.

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R. & Matgen, P. (2007) Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrology and Earth System Sciences*, 11(2), 739–752. https://doi.org/10.5194/hess-11-739-2007

Pappenberger, F., Matgen, P., Beven, K.J., Henry, J.B., Pfister, L. & de Fraipont, P. (2006) Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources*, 29(10), 1430–1449. https://doi.org/10.1016/j.advwatres.2005.11.012

Schoorl, J.M., Sonneveld, M.P.W. & Veldkamp, A. (2000) Three-dimensional landscape process modelling: The effect of DEM resolution. *Earth Surface Processes and Landforms*, 25(9), 1025–1034. https://doi.org/10.1002/1096-9837(200008)25:9%3C1025::Aid-Esp116%3E3.3.Co;2-Q

Skinner, C.J., Coulthard, T.J., Schwanghart, W., Van De Wiel, M.J. & Hancock, G. (2018) Global sensitivity analysis of parameter uncertainty in landscape evolution models. *Geoscientific Model Development*, 11(12), 4873–4888. https://doi.org/10.5194/gmd-11-4873-2018

Skinner, C.J., Peleg, N., Quinn, N., Coulthard, T.J., Molnar, P. & Freer, J. (2020) The impact of different rainfall products on landscape modelling simulations. *Earth Surface Processes and Landforms*, 45(11), 2512–2523. https://doi.org/10.1002/esp.4894

Temme, A.J.A.M., Baartman, J.E.M. & Schoorl, J.M. (2009) Can uncertain landscape evolution models discriminate between landscape responses to stable and changing future climate? A millennial-scale test. *Global and Planetary Change*, 69(1–2), 48–58. https://doi.org/10.1016/j.gloplacha.2009.08.001

Temme, A.J.A.M., Claessens, L., Veldkamp, A. & Schoorl, J.M. (2011) Evaluating choices in multi-process landscape evolution models. *Geomorphology*, 125(2), 271–281. https://doi.org/10.1016/j.geomorph.2010.10.007

Teweldebrhan, A.T., Burkhart, J.F. & Schuler, T.V. (2018) Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches. *Hydrology and Earth System Sciences*, 22(9), 5021–5039. https://doi.org/10.5194/hess-22-5021-2018

Tucker, G.E. & Hancock, G.R. (2010) Modelling landscape evolution. *Earth Surface Processes and Landforms*, 35(1), 28–50. https://doi.org/10.1002/esp.1952

Tucker, G.E. & Slingerland, R.L. (1994) Erosional dynamics, flexural isostasy, and long-lived escarpments - a numerical modeling study. *Journal of Geophysical Research - Solid Earth*, 99(B6), 12229–12243. https://doi.org/10.1029/94jb00320

Van De Wiel, M.J., Coulthard, T.J., Macklin, M.G. & Lewin, J. (2007) Embedding reach-scale fluvial dynamics within the CAESAR cellular automaton landscape evolution model. *Geomorphology*, 90(3–4), 283–301. https://doi.org/10.1016/j.geomorph.2006.10.024

Van de Wiel, M.J., Coulthard, T.J., Macklin, M.G. & Lewin, J. (2011) Modelling the response of river systems to environmental change: Progress, problems and prospects for palaeo-environmental reconstructions. *Earth-Science Reviews*, 104(1–3), 167–185. https://doi.org/10.1016/j.earscirev.2010.10.004

Van Hoey, S., Nopens, I., van der Kwast, J. & Seuntjens, P. (2015) Dynamic identifiability analysis-based model structure evaluation considering rating curve uncertainty. *Journal of Hydrologic Engineering*, 20(5), 04014072. https://doi.org/10.1061/(Asce)He.1943-5584.0000995

Verstraeten, G. & Poesen, J. (2001) Variability of dry sediment bulk density between and within retention ponds and its impact on the calculation of sediment yields. *Earth Surface Processes and Landforms*, 26(4), 375–394. https://doi.org/10.1002/esp.186

Vrugt, J.A., Clark, M.P., Diks, C.G.H., Duan, Q. & Robinson, B.A. (2006) Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters*, 33, L19817. https://doi.org/10.1029/2006gl027126

Walling, D. (1977) Assessing the accuracy of suspended sediment rating curves for a small basin. *Water Resources Research*, 13(3), 531–538. https://doi.org/10.1029/WR013i003p00531

Warburton, J. (2010) Sediment transfer in steep upland catchments (northern England, UK): landform and sediment source coupling. In: *Landform-Structure, Evolution, Process Control*. Berlin, Heidelberg: Springer, pp. 165–183.

Warburton, J., Milledge, D. & Johnson, R. (2008) Assessment of shallow landslide activity following the January 2005 storm, Northern Cumbria. *Cumberland Geological Society Proceedings*, 7, 263–283.

Welsh, K.E., Dearing, J.A., Chiverrell, R.C. & Coulthard, T.J. (2009) Testing a cellular modelling approach to simulating late-Holocene sediment and water transfer from catchment to lake in the French Alps since 1826. *Holocene*, 19(5), 785–798. https://doi.org/10.1177/0959683609105303

Westerberg, I., Guerrero, J.L., Seibert, J., Beven, K.J. & Halldin, S. (2011a) Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25(4), 603–613. https://doi.org/10.1002/hyp.7848

Westerberg, I.K., Guerrero, J.L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S. *et al.* (2011b) Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

Wilcock, P.R. & Crowe, J.C. (2003) Surface-based transport model for mixed-size sediment. *Journal of Hydraulic Engineering*, 129(2), 120–128. https://doi.org/10.1061/(Asce)0733-9429(2003)129:2(120)

Wilson, P. (1993) Plowing-boulder characteristics and associated soil properties in the Lake District and Southern Scotland. *Scottish Geographical Magazine*, 109(1), 18–26. https://doi.org/10.1080/00369229318736872

Wilson, P. (2010). *Lake District Mountain Landforms*. Lancaster: Scotforth Books.

Winsemius, H.C., Schaefli, B., Montanari, A. & Savenije, H.H.G. (2009) On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(12), W12422. https://doi.org/10.1029/2009wr007706

Wong, J.S., Freer, J.E., Bates, P.D., Sear, D.A. & Stephens, E.M. (2015) Sensitivity of a hydraulic model to channel erosion uncertainty during extreme flooding. *Hydrological Processes*, 29(2), 261–279. https://doi.org/10.1002/hyp.10148

Ziliani, L., Surian, N., Coulthard, T.J. & Tarantola, S. (2013) Reduced-complexity modeling of braided rivers: Assessing model performance by sensitivity analysis, calibration, and validation. *Journal of Geophysical Research - Earth Surface*, 118(4), 2243–2262. https://doi.org/10.1002/jgrf.20154

How to cite this article: Wong, J.S., Freer, J.E., Bates, P.D., Warburton, J. and Coulthard, T.J. (2021) Assessing the hydrological and geomorphic behaviour of a landscape evolution model within a limits-of-acceptability uncertainty analysis framework. *Earth Surface Processes and Landforms*, 1–23. Available from: https://doi.org/10.1002/esp.5140

APPENDIX

TABLE A1 Summary statistics of the DEM elevation at 50 m, 100 m and 200 m resolutions, respectively

		Percentile				
Elevation (m)		**5th**	**25th**	**50th**	**75th**	**95th**
River Derwent						
Resolution	50 m	74.0	177.5	308.4	474.7	672.9
	100 m	74.0	177.5	303.3	469.3	667.4
	200 m	74.0	177.5	301.4	467.3	668.7
Cocker						
Resolution	50 m	82.0	132.2	262.4	425.1	626.3
	100 m	81.3	131.8	260.2	421.9	623.2
	200 m	81.4	130.8	257.5	418.4	623.7

TABLE A2 Summary statistics of the slope at 50 m, 100 m and 200 m resolutions, respectively

		Percentile				
Slope (°)		**5th**	**25th**	**50th**	**75th**	**95th**
River Derwent						
Resolution	50 m	0.4	4.6	11.7	21.6	33.7
	100 m	0.4	4.3	11.0	20.2	31.6
	200 m	0.5	4.2	10.3	18.1	27.9
Cocker						
Resolution	50 m	0.8	5.6	14.0	24.2	35.6
	100 m	0.9	5.5	13.2	22.8	33.9
	200 m	1.1	5.4	11.9	20.4	30.6

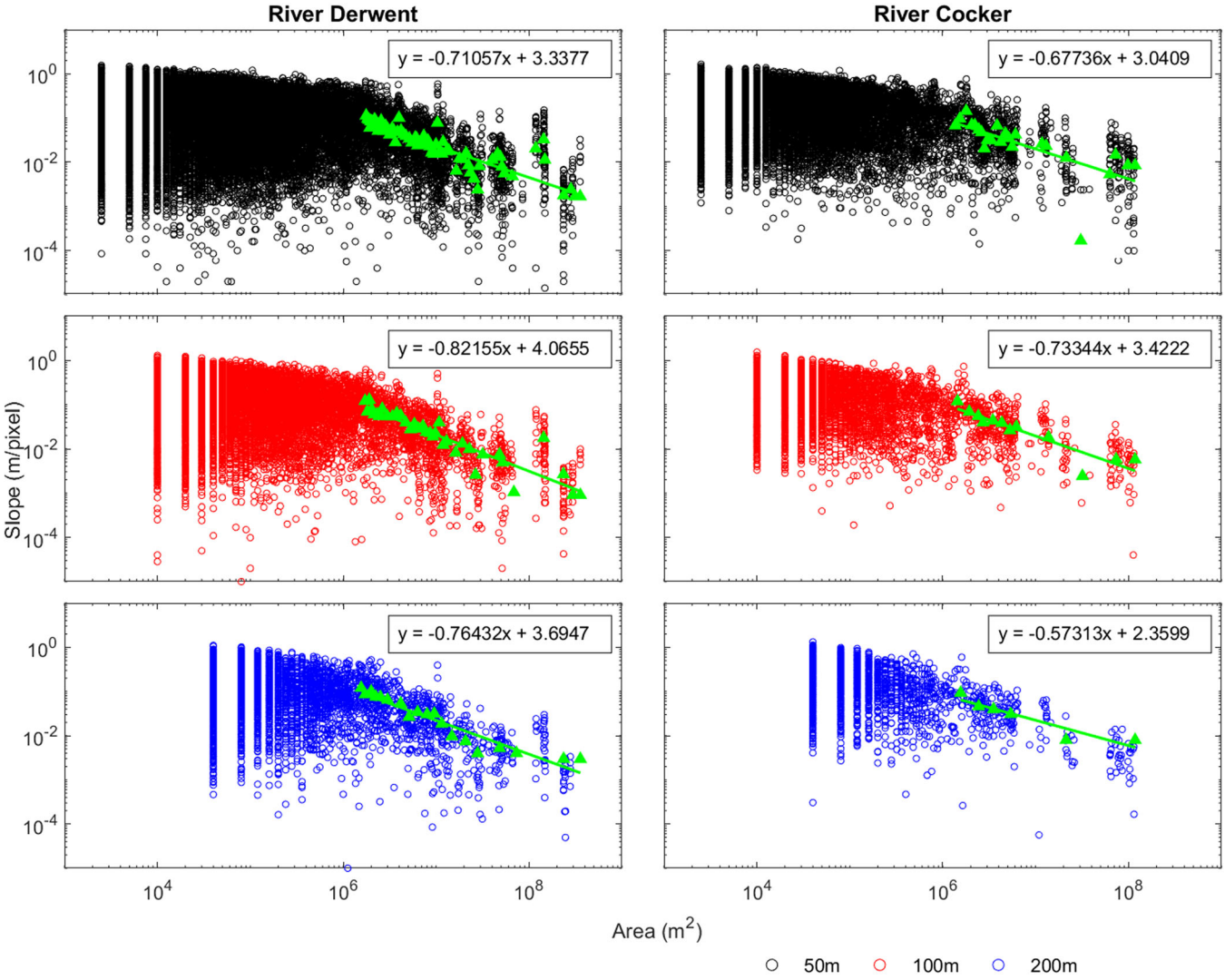

FIGURE A1 Area–slope relationship estimated on the River Derwent (left) and the River Cocker (right) based on DEM of 50 m (black hollow circle), 100 m (red hollow circle) and 200 m (blue hollow circle) resolutions. Green solid triangles represent the mean values estimated on 50 circles, whereas green solid lines are the regression line fitted on those green solid triangles

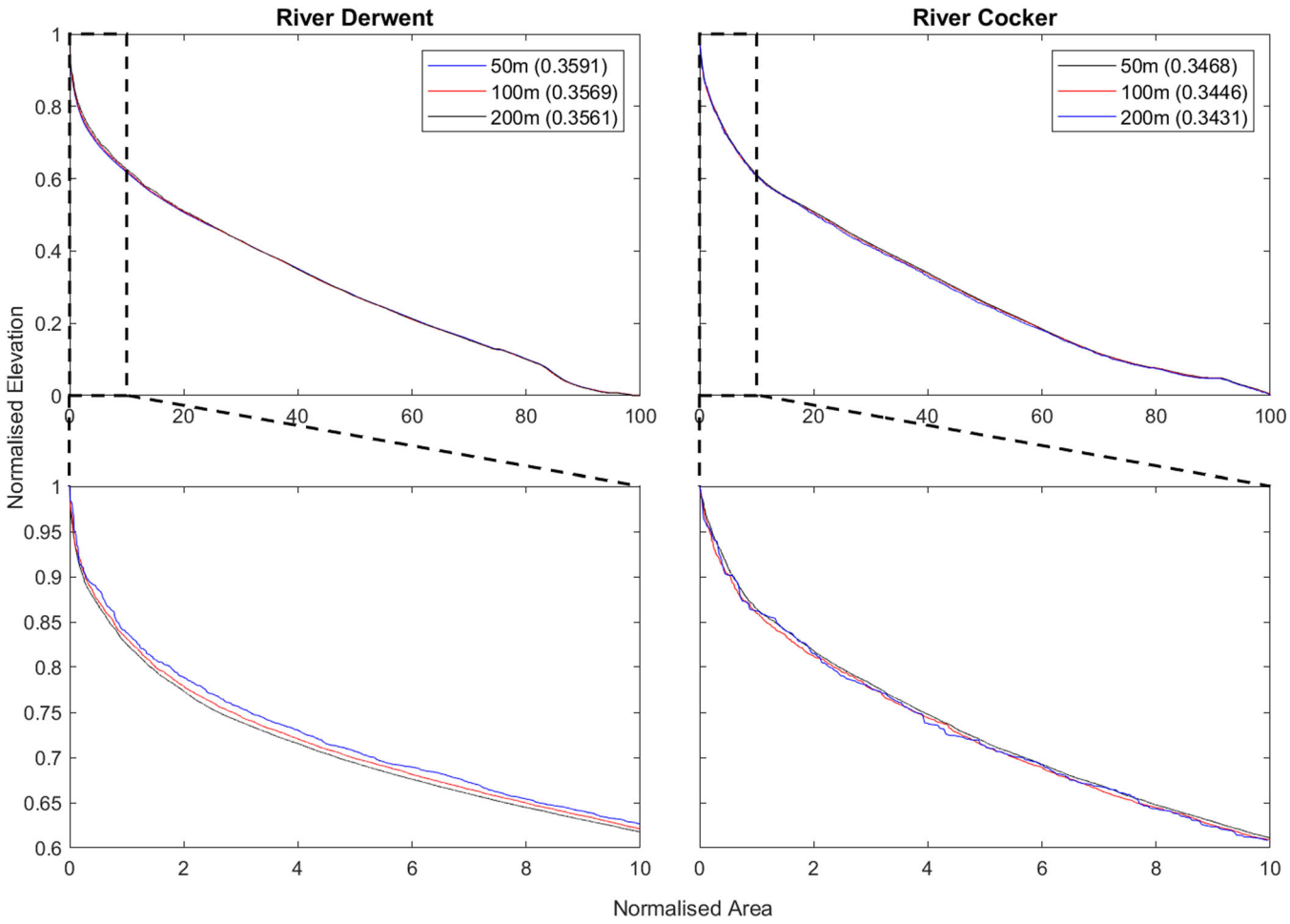

FIGURE A2 Hypsometric curve and integral estimated on the River Derwent (left) and the River cocker (right) based on DEM of 50 m (black), 100 m (red) and 200 m (blue) resolutions. The values in brackets in the legend indicate the hypsometric integrals